Emerging Technologies and Technology Forecasting

Braydon Herrmann

First Edition, 2012

ISBN 978-81-323-1476-9

Published by:
College Publishing House
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

4G

4G stands for the fourth generation of cellular wireless standards. It is a successor to 3G and 2G families of standards. Speed requirements for 4G service set the peak download speed at 100 Mbit/s for high mobility communication (such as from trains and cars) and 1 Gbit/s for low mobility communication (such as pedestrians and stationary users).

A 4G system is expected to provide a comprehensive and secure all-IP based mobile broadband solution to smartphones, laptop computer wireless modems and other mobile devices. Facilities such as ultra-broadband Internet access, IP telephony, gaming services, and streamed multimedia may be provided to users.

Pre-4G technologies such as mobile WiMAX and first-release 3G Long term evolution (LTE) have been on the market since 2006 and 2009 respectively, and are often branded as 4G. The current versions of these technologies did not fulfill the original ITU-R requirements of data rates approximately up to 1 Gbit/s for 4G systems. Marketing materials use 4G as a description for Mobile-WiMAX and LTE in their current forms.

IMT-Advanced compliant versions of the above two standards are under development and called “LTE Advanced” and “WirelessMAN-Advanced” respectively. ITU has decided that “LTE Advanced” and “WirelessMAN-Advanced” should be accorded the official designation of IMT-Advanced. On December 6, 2010, ITU announced that current versions of LTE, WiMax and other evolved 3G technologies that do not fulfill "IMT-Advanced" requirements could be considered "4G", provided they represent forerunners to IMT-Advanced and "a substantial level of improvement in performance and capabilities with respect to the initial third generation systems now deployed."

In all suggestions for 4G, the CDMA spread spectrum radio technology used in 3G systems and IS-95 is abandoned and replaced by OFDMA and other frequency-domain equalization schemes. This is combined with MIMO (Multiple In Multiple Out), e.g., multiple antennas, dynamic channel allocation and channel-dependent scheduling.

Background

The nomenclature of the generations generally refers to a change in the fundamental nature of the service, non-backwards compatible transmission technology, and new frequency bands. New generations have appeared about every ten years since the first

move from 1981 analog (1G) to digital (2G) transmission in 1992. This was followed, in 2001, by 3G multi-media support, spread spectrum transmission and at least 200 kbit/s, in 2011 expected to be followed by 4G, which refers to all-IP packet-switched networks, mobile ultra-broadband (gigabit speed) access and multi-carrier transmission.

The fastest 3G based standard in the WCDMA family is the HSPA+ standard, which was commercially available in 2009 and offers 28 Mbit/s downstreams without MIMO, i.e. only with one antenna (it would offer 56 Mbit/s with 2x2 MIMO), and 22 Mbit/s upstreams. The fastest 3G based standard in the CDMA2000 family is the EV-DO Rev. B, which was available in 2010 and offers 15.67 Mbit/s downstreams.

In mid 1990s, the ITU-R organization specified the IMT-2000 specifications for what standards that should be considered 3G systems. However, the cell phone market only brands some of the IMT-2000 standards as 3G (e.g. WCDMA and CDMA2000), but not all (3GPP EDGE, DECT and mobile-WiMAX all fulfil the IMT-2000 requirements and are formally accepted as 3G standards, but are typically not branded as 3G). In 2008, ITU-R specified the **IMT-Advanced** (*International Mobile Telecommunications Advanced*) requirements for 4G systems.

ITU Requirements and 4G wireless standards

Here we uses 4G to refer to **IMT-Advanced** (*International Mobile Telecommunications Advanced*), as defined by ITU-R. An IMT-Advanced cellular system must fulfil the following requirements:

- Based on an all-IP packet switched network.
- Peak data rates of up to approximately 100 Mbit/s for high mobility such as mobile access and up to approximately 1 Gbit/s for low mobility such as nomadic/local wireless access, according to the ITU requirements.
- Dynamically share and utilize the network resources to support more simultaneous users per cell.
- Scalable channel bandwidth, between 5 and 20 MHz, optionally up to 40 MHz.
- Peak link spectral efficiency of 15 bit/s/Hz in the downlink, and 6.75 bit/s/Hz in the uplink (meaning that 1 Gbit/s in the downlink should be possible over less than 67 MHz bandwidth)
- System spectral efficiency of up to 3 bit/s/Hz/cell in the downlink and 2.25 bit/s/Hz/cell for indoor usage
- Smooth handovers across heterogeneous networks.
- Ability to offer high quality of service for next generation multimedia support.

In September 2009, the technology proposals were submitted to the International Telecommunication Union (ITU) as 4G candidates. Basically all proposals are based on two technologies:

- LTE Advanced standardized by the 3GPP
- 802.16m standardized by the IEEE (i.e. WiMAX)

Present implementations of WiMAX and LTE are largely considered a stopgap solution that will offer a considerable boost while WiMAX 2 (based on the 802.16m spec) and LTE Advanced are finalized. Both technologies aim to reach the objectives traced by the ITU, but are still far from being implemented.

The first set of 3GPP requirements on LTE Advanced was approved in June 2008. LTE Advanced will be standardized in 2010 as part of the Release 10 of the 3GPP specification. LTE Advanced will be fully built on the existing LTE specification Release 10 and not be defined as a new specification series. A summary of the technologies that have been studied as the basis for LTE Advanced is included in a technical report.

Current LTE and WiMAX implementations are considered pre-4G, as they don't fully comply with the planned requirements of 1 Gbit/s for stationary reception and 100 Mbit/s for mobile.

Confusion has often been caused by some mobile carriers who have launched products advertised as 4G but which are actually current so-called 3.9G technologies, and therefore do not follow the ITU-R defined principles for 4G standards. A common argument for branding 3.9G systems as a new generation is that they use other frequency bands than 3G technologies, they are based on a new radio-interface paradigm, and the standards are not backwards compatible with 3G but some of them are expected to be forwards compatible with future "real" 4G technologies. While the ITU has adopted recommendations for technologies that would be used for future global communications, they do not actually do the standardization or development work themselves, instead relying on the work of other standards bodies such as IEEE, The WiMAX Forum and 3GPP. Recently, ITU-R Working Party 5D approved two industry-developed technologies (LTE Advanced and WirelessMAN-Advanced) for inclusion in the ITU's International Mobile Telecommunications Advanced (IMT-Advanced program), which is focused on global communication systems that would be available several years from now. This working party's objective was not to comment on today's 4G being rolled out in the United States and in fact, the Working Party itself purposely agreed not to tie their IMT-Advanced work to the term 4G, recognizing its common use in industry already; however, the ITU's PR department ignored that agreement and used term 4G anyway when issuing their press release.

The ITU's purpose is to foster the use of communications globally. The ITU is relied upon by developing countries, for example, who want to be assured a technology is standardised and likely to be widely deployed. While the ITU has adopted recommendations for technologies that would be used for future global communications, they do not actually do the standardization or development work themselves, instead relying on the work of other standards bodies such as IEEE, The WiMAX Forum and 3GPP. While the ITU has developed recommendations on IMT-Advanced, those recommendations are not binding on ITU member countries.

4G Predecessors and candidate systems

The wireless telecommunications industry as a whole has early assumed the term 4G as a short hand way to describe those advanced cellular technologies that, among other things, are based on or employ wide channel OFDMA and SC-FDE technologies, MIMO transmission and an all-IP based architecture. Mobile-WiMAX, first release LTE, IEEE 802.20 as well as Flash-OFDM meets these early assumptions, and have been considered as 4G candidate systems, but do not yet meet the more recent ITU-R IMT-Advanced requirements.

4G candidate systems

LTE Advanced

LTE Advanced (Long-term-evolution Advanced) is a candidate for IMT-Advanced standard, formally submitted by the 3GPP organization to ITU-T in the fall 2009, and expected to be released in 2012. The target of 3GPP LTE Advanced is to reach and surpass the ITU requirements. LTE Advanced is essentially an enhancement to LTE. It is not a new technology but rather an improvement on the existing LTE network. This upgrade path makes it more cost effective for vendors to offer LTE and then upgrade to LTE Advanced which is similar to the upgrade from WCDMA to HSPA. LTE and LTE Advanced will also make use of additional spectrum and multiplexing to allow it to achieve higher data speeds. Coordinated Multi-point Transmission will also allow more system capacity to help handle the enhanced data speeds. Release 10 of LTE is expected to achieve the LTE Advanced speeds. Release 8 currently supports up to 300 Mbit/s download speeds which is still short of the IMT-Advanced standards.

Data speeds of LTE Advanced

	LTE Advanced
Peak Download	1 Gbit/s
Peak Upload	500 Mbit/s

IEEE 802.16m or WirelessMAN-Advanced

The IEEE 802.16m or WirelessMAN-Advanced evolution of 802.16e is under development, with the objective to fulfill the IMT-Advanced criteria of 1 Gbit/s for stationary reception and 100 Mbit/s for mobile reception.

4G predecessors and discontinued candidate systems

3GPP Long Term Evolution (LTE)

Telia-branded Samsung LTE modem

The pre-4G technology 3GPP Long Term Evolution (LTE) is often branded "4G", but the first LTE release does not fully comply with the IMT-Advanced requirements. LTE has a theoretical net bit rate capacity of up to 100 Mbit/s in the downlink and 50 Mbit/s in the uplink if a 20 MHz channel is used — and more if multiple-input multiple-output (MIMO), i.e. antenna arrays, are used.

The physical radio interface was at an early stage named *High Speed OFDM Packet Access* (HSOPA), now named Evolved UMTS Terrestrial Radio Access (E-UTRA). The first LTE USB dongles do not support any other radio interface.

The world's first publicly available LTE service was opened in the two Scandinavian capitals Stockholm (Ericsson system) and Oslo (a Huawei system) on 14 December 2009, and branded 4G. The user terminals were manufactured by Samsung. Currently, the only publicly available LTE service in the United States is provided by Verizon Wireless. AT&T also has an LTE service in the works.

Mobile WiMAX (IEEE 802.16e)

The Mobile WiMAX (IEEE 802.16e-2005) mobile wireless broadband access (MWBA) standard (also known as WiBro in South Korea) is sometimes branded 4G, and offers peak data rates of 128 Mbit/s downlink and 56 Mbit/s uplink over 20 MHz wide channels.

The world's first commercial mobile WiMAX service was opened by KT in Seoul, South Korea on 30 June 2006.

Sprint Nextel has begun using Mobile WiMAX, as of September 29, 2008 branded as a "4G" network even though the current version does not fulfil the IMT Advanced requirements on 4G systems.

In Russia, Belarus and Nicaragua WiMax broadband internet access is offered by a Russian company Scartel, and is also branded 4G, Yota.

UMB (formerly EV-DO Rev. C)

UMB (Ultra Mobile Broadband) was the brand name for a discontinued 4G project within the 3GPP2 standardization group to improve the CDMA2000 mobile phone standard for next generation applications and requirements. In November 2008, Qualcomm, UMB's lead sponsor, announced it was ending development of the technology, favouring LTE instead. The objective was to achieve data speeds over 275 Mbit/s downstream and over 75 Mbit/s upstream.

Flash-OFDM

At an early stage the Flash-OFDM system was expected to be further developed into a 4G standard.

iBurst and MBWA (IEEE 802.20) systems

The iBurst system (or HC-SDMA, High Capacity Spatial Division Multiple Access) was at an early stage considered as a 4G predecessor. It was later further developed into the Mobile Broadband Wireless Access (MBWA) system, also known as IEEE 802.20.

Data rate comparison

The following table shows a comparison of 4G candidate systems as well as other competing technologies.

Comparison of Mobile Internet Access methods

Standard	Family	Primary Use	Radio Tech	Downlink (Mbit/s)	Uplink (Mbit/s)	Notes
LTE	UMTS/ 4GSM	General 4G	OFDMA/ MIMO/ SC-FDMA	100 (in 20MHz bandwidth)	50 (in 20 MHz bandwidth)	LTE-Advanced update expected to offer peak rates up to 1 Gbit/s fixed speeds and 100 Mb/s to mobile users.
WiMAX	802.16	Mobile Internet	MIMO-SOFDMA	128 (in 20MHz bandwidth)	56 (in 20MHz bandwidth)	WiMAX update IEEE 802.16m expected to offer peak rates of at least 1 Gbit/s fixed speeds and 100Mbit/s to mobile users.
Flash-OFDM	Flash-OFDM	Mobile Internet mobility up to 200mph (350km/h)	Flash-OFDM	5.3 10.6 15.9	1.8 3.6 5.4	Mobile range 30km (18 miles) extended range 55 km (34 miles)
HIPERMAN	HIPERMAN	Mobile Internet	OFDM	56.9	56.9	
Wi-Fi	802.11 (11n)	Mobile Internet	OFDM/MIMO	288.9*		Antenna, RF front end enhancements and minor

						protocol timer tweaks have helped deploy long range P2P networks compromising on radial coverage, throughput and/or spectra efficiency (310km & 382km) (*can support 600 when set at 40MHz channel width).
iBurst	802.20	Mobile Internet	HC-SDMA/ TDD/MIMO	95	36	Cell Radius: 3–12 km Speed: 250km/h Spectral Efficiency: 13 bits/s/Hz/cell Spectrum Reuse Factor: "1"
EDGE Evolution	GSM	Mobile Internet	TDMA/FDD	0.2	0.2	3GPP Release 7
UMTS W-CDMA HSDPA+ HSUPA HSPA+	UMTS/ 3GSM	General 3G	CDMA/FDD CDMA/FDD/ MIMO	0.384 14.4 56	0.384 5.76 22	HSDPA widely deployed. Typical downlink rates today 2 Mbit/s, ~200 kbit/s uplink; HSPA+

						downlink up to 56 Mbit/s.
UMTS-TDD	UMTS/ 3GSM	Mobile Internet	CDMA/TDD	16	16	Reported speeds according to IPWireless using 16QAM modulation similar to HSDPA+HSUPA
1xRTT	CDMA2000	Mobile phone	CDMA	0.144	0.144	Succeeded by EV-DO for data use, but still is used for voice and as a failover for EV-DO
EV-DO 1x Rev. 0 **EV-DO 1x Rev.A** **EV-DO Rev.B**	CDMA2000	Mobile Internet	CDMA/FDD	2.45 3.1 4.9xN	0.15 1.8 1.8xN	Rev B note: N is the number of 1.25 MHz chunks of spectrum used. EV-DO is not designed for voice, and requires a fallback to 1xRTT when a voice call is placed or received.

Notes: All speeds are theoretical maximums and will vary by a number of factors, including the use of external antennae, distance from the tower and the ground speed (e.g. communications on a train may be poorer than when standing still). Usually the bandwidth is shared between several terminals. The performance of each technology is determined by a number of constraints, including the spectral efficiency of the technology, the cell sizes used, and the amount of spectrum available.

Objective and approach

Objectives assumed in the literature

4G is being developed to accommodate the quality of service (QoS) and rate requirements set by further development of existing 3G applications like mobile broadband access, Multimedia Messaging Service (MMS), video chat, mobile TV, but also new services like HDTV. 4G may allow roaming with wireless local area networks, and may interact with digital video broadcasting systems.

In the literature, the assumed or expected 4G requirements have changed during the years before IMT-Advanced was specified by the ITU-R. These are examples of objectives stated in various sources:

- A nominal data rate of 100 Mbit/s while the client physically moves at high speeds relative to the station, and 1 Gbit/s while client and station are in relatively fixed positions as defined by the ITU-R
- A data rate of at least 100 Mbit/s between any two points in the world
- Smooth handoff across heterogeneous networks
- Seamless connectivity and global roaming across multiple networks
- High quality of service for next generation multimedia support (real time audio, high speed data, HDTV video content, mobile TV, etc.)
- Interoperability with existing wireless standards
- An all IP, packet switched network
- IP-based femtocells (home nodes connected to fixed Internet broadband infrastructure)

Approaches

Principal technologies

- Physical layer transmission techniques are as follows:
    - MIMO: To attain ultra high spectral efficiency by means of spatial processing including multi-antenna and multi-user MIMO
    - *Frequency-domain-equalization*, for example *Multi-carrier modulation (OFDM) in the downlink or* single-carrier frequency-domain-equalization *(SC-FDE) in the uplink: To exploit the frequency selective channel property without complex equalization.*
    - Frequency-domain statistical multiplexing, for example (OFDMA) or (Single-carrier FDMA) (SC-FDMA, a.k.a. Linearly precoded OFDMA, LP-OFDMA) in the uplink: Variable bit rate by assigning different sub-channels to different users based on the channel conditions
    - Turbo principle error-correcting codes: To minimize the required SNR at the reception side
- Channel-dependent scheduling: To utilize the time-varying channel.
- Link adaptation: Adaptive modulation and error-correcting codes

- Relaying, including fixed relay networks (FRNs), and the cooperative relaying concept, known as multi-mode protocol

4G features assumed in early literature

The 4G system was originally envisioned by the Defense Advanced Research Projects Agency (DARPA). The DARPA selected the distributed architecture, end-to-end Internet protocol (IP), and believed at an early stage in peer-to-peer networking in which every mobile device would be both a transceiver and a router for other devices in the network eliminating the spoke-and-hub weakness of 2G and 3G cellular systems. Since the 2.5G GPRS system, cellular systems have provided dual infrastructures: packet switched nodes for data services, and circuit switched nodes for voice calls. In 4G systems, the circuit-switched infrastructure is abandoned, and only a packet-switched network is provided, while 2.5G and 3G systems require both packet-switched and circuit-switched network nodes, i.e. two infrastructures in parallel. This means that in 4G, traditional voice calls are replaced by IP telephony.

Cellular systems such as 4G allow seamless mobility; thus a file transfer is not interrupted in case a terminal moves from one cell (one base station coverage area) to another, but handover is carried out. The terminal also keeps the same IP address while moving, meaning that a mobile server is reachable as long as it is within the coverage area of any server. In 4G systems this mobility is provided by the mobile IP protocol, part of IP version 6, while in earlier cellular generations it was only provided by physical layer and datalink layer protocols. In addition to seamless mobility, 4G provides flexible interoperability of the various kinds of existing wireless networks, such as satellite, cellular wirelss, WLAN, PAN and systems for acessing fixed wireless networks.

While maintaining seamless mobility, 4G will offer very high data rates with expectations of 100 Mbit/s wireless service. The increased bandwidth and higher data transmission rates will allow 4G users the ability to utilize high definition video and the video conferencing features of mobile devices attached to a 4G network. The 4G wireless system is expected to provide a comprehensive IP solution where multimedia applications and services can be delivered to the user on an ‘Anytime, Anywhere' basis with a satisfactory high data rate, premium quality and high security.

4G is described as MAGIC — Mobile multimedia, Anytime anywhere, Global mobility support, Integrated wireless solution, and Customized personal service.

Some key features (primarily from users' points of view) of 4G mobile networks are as follows:

- High usability: anytime, anywhere, and with any technology
- Support for multimedia services at low transmission cost
- Personalization
- Integrated services

Some candidate systems suggest having an open Internet platform.

Components

Access schemes

As the wireless standards evolved, the access techniques used also exhibited increase in efficiency, capacity and scalability. The first generation wireless standards used plain TDMA and FDMA. In the wireless channels, TDMA proved to be less efficient in handling the high data rate channels as it requires large guard periods to alleviate the multipath impact. Similarly, FDMA consumed more bandwidth for guard to avoid inter carrier interference. So in second generation systems, one set of standard used the combination of FDMA and TDMA and the other set introduced an access scheme called CDMA. Usage of CDMA increased the system capacity, but as a theoretical drawback placed a soft limit on it rather than the hard limit (i.e. a CDMA network setup does not inherently reject new clients when it approaches its limits, resulting in a denial of service to all clients when the network overloads; though this outcome is avoided in practical implementations by admission control of circuit switched or fixed bitrate communication services). Data rate is also increased as this access scheme (providing the network is not reaching its capacity) is efficient enough to handle the multipath channel. This enabled the third generation systems, such as IS-2000, UMTS, HSXPA, 1xEV-DO, TD-CDMA and TD-SCDMA, to use CDMA as the access scheme. However, the issue with CDMA is that it suffers from poor spectral flexibility and computationally intensive time-domain equalization (high number of multiplications per second) for wideband channels.

Recently, new access schemes like Orthogonal FDMA (OFDMA), Single Carrier FDMA (SC-FDMA), Interleaved FDMA and Multi-carrier CDMA (MC-CDMA) are gaining more importance for the next generation systems. These are based on efficient FFT algorithms and frequency domain equalization, resulting in a lower number of multiplications per second. They also make it possible to control the bandwidth and form the spectrum in a flexible way. However, they require advanced dynamic channel allocation and traffic adaptive scheduling.

WiMax is using OFDMA in the downlink and in the uplink. For the next generation UMTS, OFDMA is used for the downlink. By contrast, IFDMA is being considered for the uplink since OFDMA contributes more to the PAPR related issues and results in nonlinear operation of amplifiers. IFDMA provides less power fluctuation and thus avoids amplifier issues. Similarly, MC-CDMA is in the proposal for the IEEE 802.20 standard. These access schemes offer the same efficiencies as older technologies like CDMA. Apart from this, scalability and higher data rates can be achieved.

The other important advantage of the above mentioned access techniques is that they require less complexity for equalization at the receiver. This is an added advantage especially in the MIMO environments since the spatial multiplexing transmission of MIMO systems inherently requires high complexity equalization at the receiver.

In addition to improvements in these multiplexing systems, improved modulation techniques are being used. Whereas earlier standards largely used Phase-shift keying, more efficient systems such as 64QAM are being proposed for use with the 3GPP Long Term Evolution standards.

IPv6 support

Unlike 3G, which is based on two parallel infrastructures consisting of circuit switched and packet switched network nodes respectively, 4G will be based on packet switching *only*. This will require low-latency data transmission.

By the time that 4G is deployed, the process of IPv4 address exhaustion is expected to be in its final stages. Therefore, in the context of 4G, IPv6 support is essential in order to support a large number of wireless-enabled devices. By increasing the number of IP addresses, IPv6 removes the need for network address translation (NAT), a method of sharing a limited number of addresses among a larger group of devices, although NAT will still be required to communicate with devices that are on existing IPv4 networks.

As of June 2009, Verizon has posted specifications that require any 4G devices on its network to support IPv6.

Advanced antenna systems

The performance of radio communications depends on an antenna system, termed smart or intelligent antenna. Recently, multiple antenna technologies are emerging to achieve the goal of 4G systems such as high rate, high reliability, and long range communications. In the early 1990s, to cater for the growing data rate needs of data communication, many transmission schemes were proposed. One technology, spatial multiplexing, gained importance for its bandwidth conservation and power efficiency. Spatial multiplexing involves deploying multiple antennas at the transmitter and at the receiver. Independent streams can then be transmitted simultaneously from all the antennas. This technology, called MIMO (as a branch of intelligent antenna), multiplies the base data rate by (the smaller of) the number of transmit antennas or the number of receive antennas. Apart from this, the reliability in transmitting high speed data in the fading channel can be improved by using more antennas at the transmitter or at the receiver. This is called *transmit* or *receive diversity*. Both transmit/receive diversity and transmit spatial multiplexing are categorized into the space-time coding techniques, which does not necessarily require the channel knowledge at the transmitter. The other category is closed-loop multiple antenna technologies, which require channel knowledge at the transmitter.

Software-defined radio (SDR)

SDR is one form of open wireless architecture (OWA). Since 4G is a collection of wireless standards, the final form of a 4G device will constitute various standards. This

can be efficiently realized using SDR technology, which is categorized to the area of the radio convergence.

History of 4G and pre-4G technologies

- In 2002, the strategic vision for 4G—which ITU designated as IMT-Advanced—was laid out.
- In 2005, OFDMA transmission technology is chosen as candidate for the HSOPA downlink, later renamed 3GPP Long Term Evolution (LTE) air interface E-UTRA.
- In November 2005, KT demonstrated mobile WiMAX service in Busan, South Korea.
- In June 2006, KT started the world's first commercial mobile WiMAX service in Seoul, South Korea.
- In mid-2006, Sprint Nextel announced that it would invest about US$5 billion in a WiMAX technology buildout over the next few years ($5.45 billion in real terms). Since that time Sprint has faced many setbacks, that have resulted in steep quarterly losses. On May 7, 2008, Sprint, Imagine, Google, Intel, Comcast, Bright House, and Time Warner announced a pooling of an average of 120 MHz of spectrum; Sprint merged its Xohm WiMAX division with Clearwire to form a company which will take the name "Clear".
- In February 2007, the Japanese company NTT DoCoMo tested a 4G communication system prototype with 4x4 MIMO called VSF-OFCDM at 100 Mbit/s while moving, and 1 Gbit/s while stationary. NTT DoCoMo completed a trial in which they reached a maximum packet transmission rate of approximately 5 Gbit/s in the downlink with 12x12 MIMO using a 100 MHz frequency bandwidth while moving at 10 km/h, and is planning on releasing the first commercial network in 2010.
- In September 2007, NTT Docomo demonstrated e-UTRA data rates of 200 Mbit/s with power consumption below 100 mW during the test.
- In January 2008, a U.S. Federal Communications Commission (FCC) spectrum auction for the 700 MHz former analog TV frequencies began. As a result, the biggest share of the spectrum went to Verizon Wireless and the next biggest to AT&T. Both of these companies have stated their intention of supporting LTE.
- In January 2008, EU commissioner Viviane Reding suggested re-allocation of 500–800 MHz spectrum for wireless communication, including WiMAX.
- February 15, 2008 - Skyworks Solutions released a front-end module for e-UTRAN.
- In April 2008, LG and Nortel demonstrated e-UTRA data rates of 50 Mbit/s while travelling at 110 km/h.
- In 2008, ITU-R established the detailed performance requirements of IMT-Advanced, by issuing a Circular Letter calling for candidate Radio Access Technologies (RATs) for IMT-Advanced.
- April 2008, just after receiving the circular letter, the 3GPP organized a workshop on IMT-Advanced where it was decided that LTE Advanced, an evolution of

current LTE standard, will meet or even exceed IMT-Advanced requirements following the ITU-R agenda.
- On 3 March 2009, Lithuania's LRTC announcing the first operational "4G" mobile WiMAX network in Baltic states.
- In December 2009, Sprint began advertising "4G" service in selected cities in the United States, despite average download speeds of only 3–6 Mbit/s with peak speeds of 10 Mbit/s (not available in all markets).
- On December 14, 2009, the first commercial LTE deployment was in the Scandinavian capitals Stockholm and Oslo by the Swedish-Finnish network operator TeliaSonera and its Norwegian brandname NetCom (Norway). TeliaSonera branded the network "4G". The modem devices on offer were manufactured by Samsung (dongle GT-B3710), and the network infrastructure created by Huawei (in Oslo) and Ericsson (in Stockholm). TeliaSonera plans to roll out nationwide LTE across Sweden, Norway and Finland. TeliaSonera used spectral bandwidth of 10 MHz, and single-in-single-out, which should provide physical layer net bitrates of up to 50 Mbit/s downlink and 25 Mbit/s in the uplink. Introductory tests showed a TCP throughput of 42.8 Mbit/s downlink and 5.3 Mbit/s uplink in Stockholm.
- On 25 February 2010, Estonia's EMT opened LTE "4G" network working in test regime.
- On 4 June 2010, Sprint Nextel released the first 4G Smartphone, the HTC Evo 4G.
- On July 2010, Uzbekistan's MTS deployed LTE in Tashkent.
- On 25 August 2010, Latvia's LMT opened LTE "4G" network working in test regime 50% of territory.
- On 6 December 2010, at the ITU World Radiocommunication Seminar 2010, the ITU stated that LTE, WiMax and similar "evolved 3G technologies" could be considered "4G".
- On 12 December 2010, VivaCell-MTS launches in Armenia 4G/LTE commercial test network with a live demo conducted in Yerevan.

Deployment plans

In May 2005, Digiweb, an Irish fixed and wireless broadband company, announced that they had received a mobile communications license from the Irish Telecoms regulator, ComReg. This service will be issued the mobile code *088* in Ireland and will be used for the provision of 4G Mobile communications. Digiweb launched a mobile broadband network using FLASH-OFDM technology at 872 MHz.

On September 20, 2007, Verizon Wireless announced plans for a joint effort with the Vodafone Group to transition its networks to the 4G standard LTE. On December 9, 2008, Verizon Wireless announced their intentions to build and begin to roll out an LTE network by the end of 2009. Since then, Verizon Wireless has said that they will start their rollout by the end of 2010.

On July 7, 2008, South Korea announced plans to spend 60 billion won, or US$58,000,000, on developing 4G and even 5G technologies, with the goal of having the highest mobile phone market share by 2012, and the hope of an international standard.

Telus and Bell Canada, the major Canadian cdmaOne and EV-DO carriers, have announced that they will be cooperating towards building a fourth generation (4G) LTE wireless broadband network in Canada. As a transitional measure, they are implementing 3G UMTS that went live in November 2009.

Sprint offers a 3G/4G connection plan, currently available in select cities in the United States. It delivers rates up to 10 Mbit/s.

In the United Kingdom, Telefónica O_2 is to use Slough as a guinea pig in testing the 4G network and has called upon Huawei to install LTE technology in six masts across the town to allow people to talk to each other via HD video conferencing and play PlayStation games while on the move.

Verizon Wireless has announced that it plans to augment its CDMA2000-based EV-DO 3G network in the United States with LTE. AT&T, along with Verizon Wireless, has chosen to migrate toward LTE from 2G/GSM and 3G/HSPA by 2011.

Sprint Nextel has deployed WiMAX technology which it has labeled 4G as of October 2008. It is currently deploying to additional markets and is the first US carrier to offer a WiMAX phone.

The U.S. FCC is exploring the possibility of deployment and operation of a nationwide 4G public safety network which would allow first responders to seamlessly communicate between agencies and across geographies, regardless of devices. In June 2010 the FCC released a comprehensive white paper which indicates that the 10 MHz of dedicated spectrum currently allocated from the 700 MHz spectrum for public safety will provide adequate capacity and performance necessary for normal communications as well as serious emergency situations.

TeliaSonera started deploying LTE (branded "4G") in Stockholm and Oslo November 2009 (as seen above), and in several Swedish, Norwegian, and Finnish cities during 2010. In June 2010, Swedish television companies used 4G to broadcast live television from the Swedish Crown Princess' Royal Wedding.

Beyond 4G research

A major issue in 4G systems is to make the high bit rates available in a larger portion of the cell, especially to users in an exposed position in between several basestations. In current research, this issue is addressed by macro-diversity techniques, also known as group cooperative relay, and also by beam-division multiple access.

Pervasive networks are an amorphous and at present entirely hypothetical concept where the user can be simultaneously connected to several wireless access technologies and can seamlessly move between them. These access technologies can be Wi-Fi, UMTS, EDGE, or any other future access technology. Included in this concept is also smart-radio (also known as cognitive radio technology) to efficiently manage spectrum use and transmission power as well as the use of mesh routing protocols to create a pervasive network.

Safaricom, the largest telecommunication company in East& Central Africa by both numbers of up to 17M subscribers with revenues topping US$1B& over US$300M in profits as per the 2009/2010 financial year, began it's muilti-million dollar setup of 4G network in October 2010 after the now retired& Kenya Tourist Board Chairman, Michael Joseph, regarded their 3G network as a white elephant i.e it failed to perform to expectations.

Huawei was given the contract the network is set to go fully commercial by the end of Q1 of 2011.

Chapter 2

Genetic Engineering

Genetic engineering, also called **genetic modification**, is the direct human manipulation of an organism's genetic material in a way that does not occur under natural conditions. It involves the use of recombinant DNA techniques, but does not include traditional animal and plant breeding or mutagenesis. Any organism that is generated using these techniques is considered to be a genetically modified organism. The first organisms genetically engineered were bacteria in 1973 and then mice in 1974. Insulin producing bacteria were commercialized in 1982 and genetically modified food has been sold since 1994.

The most common form of genetic engineering involves the insertion of new genetic material at an unspecified location in the host genome. This is accomplished by isolating and copying the genetic material of interest, generating a construct containing all the genetic elements for correct expression, and then inserting this construct into the host organism. Other forms of genetic engineering include gene targeting and knocking out specific genes via engineered nucleases such as zinc finger nucleases or engineered homing endonucleases.

Genetic engineering techniques have been applied in numerous fields including research, biotechnology, and medicine. Medicines such as insulin and human growth hormone are now produced in bacteria, experimental mice such as the oncomouse and the knockout mouse are being used for research purposes and insect resistant and/or herbicide tolerant crops have been commercialized. Genetically engineered plants and animals capable of producing biotechnology drugs more cheaply than current methods (called pharming) are also being developed and in 2009 the FDA approved the sale of the pharmaceutical protein antithrombin produced in the milk of genetically engineered goats.

Definition

Genetic engineering alters the genetic makeup of an organism using techniques that introduce heritable material prepared outside the organism either directly into the host or into a cell that is then fused or hybridized with the host. This involves using recombinant nucleic acid (DNA or RNA) techniques to form new combinations of heritable genetic material followed by the incorporation of that material either indirectly through a vector system or directly through micro-injection, macro-injection and micro-encapsulation techniques. Genetic engineering does not include traditional animal and plant breeding, in vitro fertilisation, induction of polyploidy, mutagenesis and cell fusion techniques that do

not use recombinant nucleic acids or a genetically modified organism in the process. Cloning and stem cell research, although not considered genetic engineering, are closely related and genetic engineering can be used within them. Synthetic biology is an emerging discipline that takes genetic engineering a step further by introducing artificially synthesized genetic material from raw materials into an organism.

If genetic material from another species is added to the host, the resulting organism is called transgenic. If genetic material from the same species or a species that can naturally breed with the host is used the resulting organism is called cisgenic. Genetic engineering can also be used to remove genetic material from the target organism, creating a knock out organism. In Europe genetic modification is synonymous with genetic engineering while within the United States of America it can also refer to conventional breeding methods.

History

Humans have altered the genomes of species for thousands of years through artificial selection and more recently mutagenesis. Genetic engineering as the direct manipulation of DNA by humans outside breeding and mutations has only existed since the 1970s. The term "genetic engineering" was first coined by Jack Williamson in his science fiction novel *Dragon's Island*, published in 1951, one year before DNA's role in heredity was confirmed by Alfred Hershey and Martha Chase, and two years before James Watson and Francis Crick showed that the DNA molecule has a double-helix structure.

In 1972 Paul Berg created the first recombinant DNA molecules by combined DNA from the monkey virus SV40 with that of the lambda virus. In 1973 Herbert Boyer and Stanley Cohen created the first transgenic organism by inserting antibiotic resistance genes into the plasmid of an *E. coli* bacterium. A year later Rudolf Jaenisch created a transgenic mouse by introducing foreign DNA into its embryo, making it the world's first transgenic animal. In 1976 Genentech, the first genetic engineering company was founded by Herbert Boyer and Robert Swanson and a year later and the company produced a human protein (somatostatin) in *E.coli*. Genentech announced the production of genetically engineered human insulin in 1978. In 1980, the U.S. Supreme Court in the Diamond v. Chakrabarty case ruled that genetically altered life could be patented. The insulin produced by bacteria, branded humulin, was approved for release by the Food and Drug Administration in 1982.

The first field trials of genetically engineered plants occurred in France and the USA in 1986, tobacco plants were engineered to be resistant to herbicides. The People's Republic of China was the first country to commercialize transgenic plants, introducing a virus-resistant tobacco in 1992. In 1994 Calgene attained approval to commercially release the Flavr Savr tomato, a tomato engineered to have a longer shelf life. In 1994, the European Union approved tobacco engineered to be resistant to the herbicide bromoxynil, making it the first genetically engineered crop commercialized in Europe. In 1995, Bt Potato was approved safe by the Environmental Protection Agency, making it the first pesticide producing crop to be approved in the USA. In 2009 11 transgenic crops were grown

commercially in 25 countries, the largest of which by area grown were the USA, Brazil, Argentina, India, Canada, China, Paraguay and South Africa.

In 2010, scientists at the J. Craig Venter Institute, announced that they had created the first synthetic bacterial genome, and added it to a cell containing no DNA. The resulting bacterium, named Synthia, was the world's first synthetic life form.

Process

Isolating the Gene

Elements of genetic engineering

First, the gene to be inserted into the genetically modified organism must be chosen and isolated. Presently, most genes transferred into plants provide protection against insects or tolerance to herbicides. In animals the majority of genes used are growth hormone genes. Once chosen the genes must be isolated. This typically involves multiplying the gene using polymerase chain reaction (PCR). If the chosen gene or the donor organism's genome has been well studied it may be present in a genetic library. If the DNA sequence is known, but no copies of the gene are available, it can be artificially synthesized. Once isolated, the gene is inserted into a bacterial plasmid.

Constructs

The gene to be inserted into the genetically modified organism must be combined with other genetic elements in order for it to work properly. The gene can also be modified at this stage for better expression or effectiveness. As well as the gene to be inserted most constructs contain a promoter and terminator region as well as a selectable marker gene. The promoter region initiates transcription of the gene and can be used to control the location and level of gene expression, while the terminator region ends transcription. The selectable marker, which in most cases confers antibiotic resistance to the organism it is expressed in, is needed to determine which cells are transformed with the new gene. The constructs are made using recombinant DNA techniques, such as restriction digests, ligations and molecular cloning.

Gene Targeting

The most common form of genetic engineering involves inserting new genetic material randomly within the host genome. Other techniques allow new genetic material to be inserted at a specific location in the host genome or generate mutations at desired genomic loci capable of knocking out endogenous genes. The technique of gene targeting uses homologous recombination to target desired changes to a specific endogenous gene. This tends to occur at a relatively low frequency in plants and animals and generally requires the use of selectable markers. The frequency of gene targeting can be greatly enhanced with the use of engineered nucleases such as zinc finger nucleases, engineered homing endonucleases, or nucleases created from TAL effectors. In addition to enhancing gene targeting, engineered nucleases can also be used to introduce mutations at endogenous genes that generate a gene knockout.

Transformation

A. tumefaciens attaching itself to a carrot cell

About 1% of bacteria are naturally able to take up foreign DNA but it can also be induced in other bacteria. Stressing the bacteria for example, with a heat shock or an electric

shock, can make the cell membrane permeable to DNA that may then incorporate into their genome or exist as extrachromosomal DNA. DNA is generally inserted into animal cells using microinjection, where it can be injected through the cells nuclear envelope directly into the nucleus or through the use of viral vectors. In plants the DNA is generally inserted using *Agrobacterium*-mediated recombination or biolistics.

In *Agrobacterium*-mediated recombination the plasmid construct must also contain T-DNA. *Agrobacterium* naturally inserts DNA from a tumor inducing plasmid into any susceptible plant's genome it infects, causing crown gall disease. The T-DNA region of this plasmid is responsible for insertion of the DNA. The genes to be inserted are cloned into a binary vector, which contains T-DNA and can be grown in both *E. Coli* and *Agrobacterium*. Once the binary vector is constructed the plasmid is transformed into *Agrobacterium* containing no plasmids and plant cells are infected. The *Agrobacterium* will then naturally insert the genetic material into the plant cells.

In biolistics particles of gold or tungsten are coated with DNA and then shot into young plant cells or plant embryos. Some genetic material will enter the cells and transform them. This method can be used on plants that are not susceptible to *Agrobacterium* infection and also allows transformation of plant plastids. Another transformation method for plant and animal cells is electroporation. Electroporation involves subjecting the plant or animal cell to an electric shock, which can make the cell membrane permeable to plasmid DNA. In some cases the electroporated cells will incorporate the DNA into their genome. Due to the damage caused to the cells and DNA the transformation efficiency of biolistics and electroporation is lower than agrobacterial mediated transformation and microinjection.

Selection

Not all the organism's cells will be transformed with the new genetic material; in most cases a selectable marker is used to differentiate transformed from untransformed cells. If a cell has been successfully transformed with the DNA it will also contain the marker gene. By growing the cells in the presence of an antibiotic or chemical that selects or marks the cells expressing that gene it is possible to separate the transgenic events from the non-transgenic. Another method of screening involves using a DNA probe that will only stick to the inserted gene. A number of strategies have been developed that can remove the selectable marker from the mature transgenic plant.

Regeneration

As often only a single cell is transformed with genetic material the organism must be regrown from that single cell. As bacteria consist of a single cell and reproduce clonally regeneration is not necessary. In plants this is accomplished through the use of tissue culture. Each plant species has different requirements for successful regeneration through tissue culture. If successful an adult plant is produced that contains the transgene in every cell. In animals it is necessary to ensure that the inserted DNA is present in the embryonic stem cells. When the offspring is produced they can be screened for the

presence of the gene. All offspring from the first generation will be heterozygous for the inserted gene and must be mated together to produce a homozygous animal.

Confirmation

Further tests using PCR, Southern Blots and Bioassays are needed to confirm that the gene is expressed and functions correctly. The organism's offspring are also tested to ensure that the trait can be inherited and that it follows a Mendelian inheritance pattern.

Applications

Genetic engineering has applications in medicine, research, industry and agriculture and can be used on a wide range of plants, animals and micro organism.

Medicine

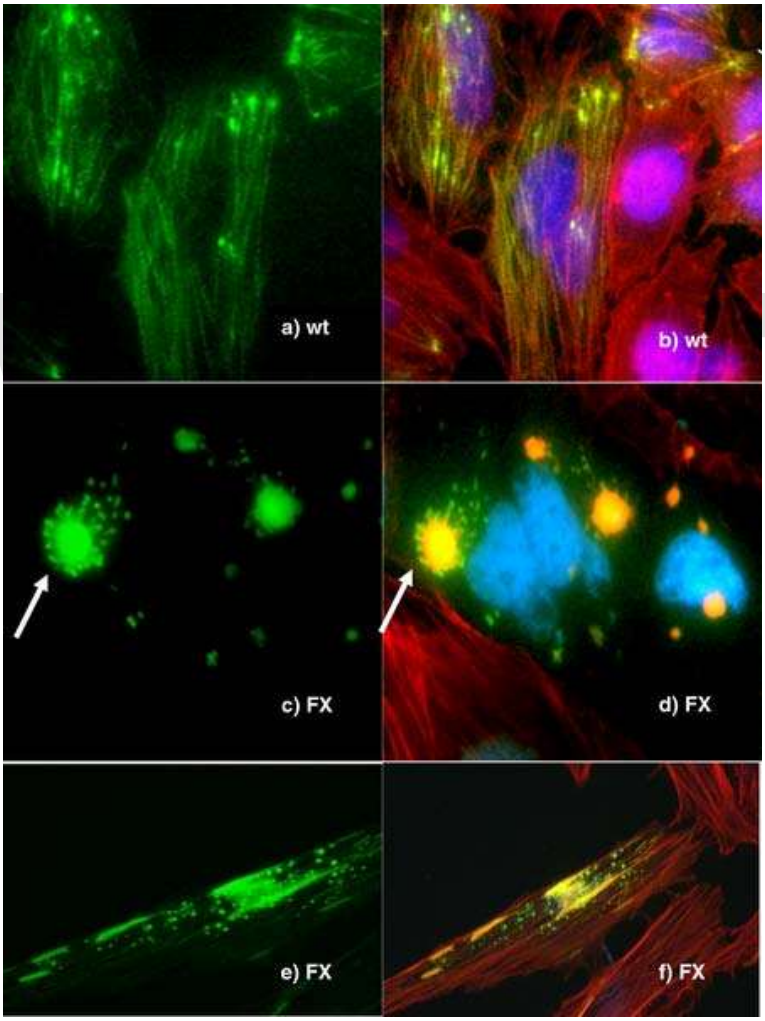
In medicine genetic engineering has been used to mass-produce insulin, human growth hormones, follistim (for treating infertility), human albumin, monoclonal antibodies, antihemophilic factors, vaccines and many other drugs. Vaccination generally involves injecting weak live, killed or inactivated forms of viruses or their toxins into the person being immunized. Genetically engineered viruses are being developed that can still confer immunity, but lack the infectious sequences. Mouse hybridomas, cells fused together to create monoclonal antibodies, have been humanised through genetic engineering to create human monoclonal antibodies.

Genetic engineering is used to create animal models of human diseases. Genetically modified mice are the most common genetically engineered animal model. They have been used to study and model cancer (the oncomouse), obesity, heart disease, diabetes, arthritis, substance abuse, anxiety, aging and Parkinson disease. Potential cures can be tested against these mouse models. Also genetically modified pigs have been bred with the aim of increasing the success of pig to human organ transplantation.

Gene therapy is the genetic engineering of humans by replacing defective human genes with functional copies. This can occur in somatic tissue or germline tissue. If the gene is inserted into the germline tissue it can be passed down to that person's descendants. Gene therapy has been used to treat patients suffering from immune deficiencies (notably Severe combined immunodeficiency) and trials have been carried out on other genetic disorders. The success of gene therapy so far has been limited and a patient (Jesse Gelsinger) has died during a clinical trial testing a new treatment. There are also ethical concerns should the technology be used not just for treatment, but for enhancement, modification or alteration of a human beings' appearance, adaptability, intelligence, character or behavior. The distinction between cure and enhancement can also be difficult to establish. Transhumanists consider the enhancement of humans desirable.

Research

Knockout mice



Human cells in which some proteins are fused with green fluorescent protein to allow them to be visualised

Genetic engineering is an important tool for natural scientists. Genes and other genetic information from a wide range of organisms are transformed into bacteria for storage and modification, creating genetically modified bacteria in the process. Bacteria are cheap, easy to grow, clonal, multiply quickly, relatively easy to transform and can be stored at -

80°C almost indefinitely. Once a gene is isolated it can be stored inside the bacteria providing an unlimited supply for research.

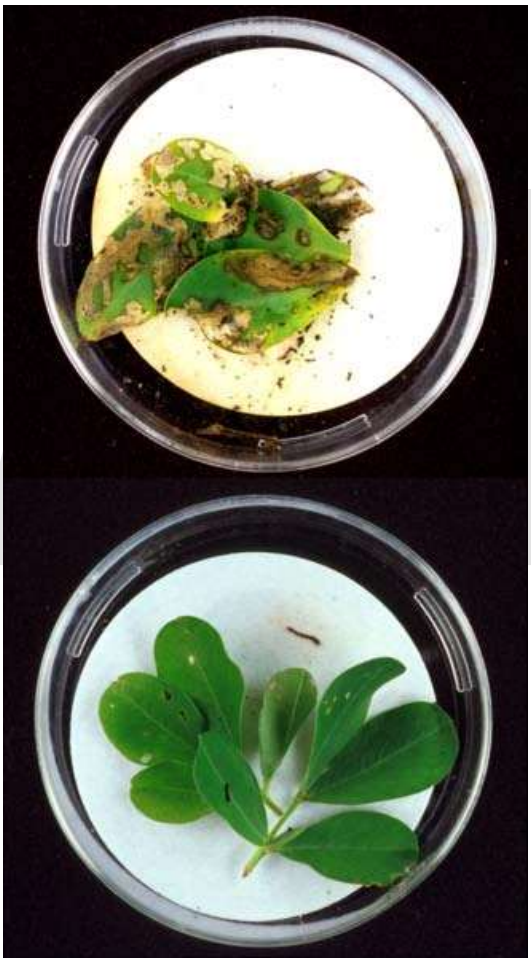
Organisms are genetically engineered to discover the functions of certain genes. This could be the effect on the phenotype of the organism, where the gene is expressed or what other genes it interacts with. These experiments generally involve loss of function, gain of function, tracking and expression.

- **Loss of function experiments**, such as in a gene knockout experiment, in which an organism is engineered to lack the activity of one or more genes. A knockout experiment involves the creation and manipulation of a DNA construct *in vitro*, which, in a simple knockout, consists of a copy of the desired gene, which has been altered such that it is non-functional. Embryonic stem cells incorporate the altered gene, which replaces the already present functional copy. These stem cells are injected into blastocysts, which are implanted into surrogate mothers. This allows the experimenter to analyze the defects caused by this mutation and thereby determine the role of particular genes. It is used especially frequently in developmental biology. Another method, useful in organisms such as Drosophila (fruit fly), is to induce mutations in a large population and then screen the progeny for the desired mutation. A similar process can be used in both plants and prokaryotes.
- **Gain of function experiments**, the logical counterpart of knockouts. These are sometimes performed in conjunction with knockout experiments to more finely establish the function of the desired gene. The process is much the same as that in knockout engineering, except that the construct is designed to increase the function of the gene, usually by providing extra copies of the gene or inducing synthesis of the protein more frequently.
- **Tracking experiments**, which seek to gain information about the localization and interaction of the desired protein. One way to do this is to replace the wild-type gene with a 'fusion' gene, which is a juxtaposition of the wild-type gene with a reporting element such as green fluorescent protein (GFP) that will allow easy visualization of the products of the genetic modification. While this is a useful technique, the manipulation can destroy the function of the gene, creating secondary effects and possibly calling into question the results of the experiment. More sophisticated techniques are now in development that can track protein products without mitigating their function, such as the addition of small sequences that will serve as binding motifs to monoclonal antibodies.
- **Expression studies** aim to discover where and when specific proteins are produced. In these experiments, the DNA sequence before the DNA that codes for a protein, known as a gene's promoter, is reintroduced into an organism with the protein coding region replaced by a reporter gene such as GFP or an enzyme that catalyzes the production of a dye. Thus the time and place where a particular protein is produced can be observed. Expression studies can be taken a step further by altering the promoter to find which pieces are crucial for the proper expression of the gene and are actually bound by transcription factor proteins; this process is known as promoter bashing.

Industrial

By engineering genes into bacterial plasmids it is possible to create a biological factory that can produce proteins and enzymes. Some genes do not work well in bacteria, so yeast, a eukaryote, can also be used. Bacteria and yeast factories have been used to produce medicines such as insulin, human growth hormone, and vaccines, supplements such as tryptophan, aid in the production of food (chymosin in cheese making) and fuels. Other applications involving genetically engineered bacteria being investigated involve making the bacteria perform tasks outside their natural cycle, such as cleaning up oil spills, carbon and other toxic waste.

Agriculture



Bt-toxins present in peanut leaves (bottom image) protect it from extensive damage caused by European corn borer larvae (top image).

One of the best-known and controversial applications of genetic engineering is the creation of genetically modified food. There are three generations of genetically modified crops. First generation crops have been commercialized and most provide protection from insects and/or resistance to herbicides. There are also fungal and virus resistant crops

developed or in development. They have been developed to make the insect and weed management of crops easier and can indirectly increase crop yield.

The second generation of genetically modified crops being developed aim to directly improve yield by improving salt, cold or drought tolerance and to increase the nutritional value of the crops. The third generation consists of pharmaceutical crops, crops that contain edible vaccines and other drugs. Some agriculturally important animals have been genetically modified with growth hormones to increase their size while others have been engineered to express drugs and other proteins in their milk.

The genetic engineering of agricultural crops can increase the growth rates and resistance to different diseases caused by pathogens and parasites. This is beneficial as it can greatly increase the production of food sources with the usage of fewer resources that would be required to host the world's growing populations. These modified crops would also reduce the usage of chemicals, such as fertilizers and pesticides, and therefore decrease the severity and frequency of the damages produced by these chemical pollution.

Ethical and safety concerns have been raised around the use of genetically modified food. A major safety concern relates to the human health implications of eating genetically modified food, in particular whether toxic or allergic reactions could occur. Gene flow into related non-transgenic crops, off target effects on beneficial organisms and the impact on biodiversity are important environmental issues. Ethical concerns involve religious issues, corporate control of the food supply, intellectual property rights and the level of labeling needed on genetically modified products.

Other uses

In materials science, a genetically modified virus has been used to construct a more environmentally friendly lithium-ion battery. Some bacteria have been genetically engineered to create black and white photographs while others have potential to be used as sensors by expressing a fluorescent protein under certain environmental conditions. Genetic engineering is also being used to create BioArt and novelty items such as blue roses, and glowing fish.

Opposition and criticism

A 2010 study of Canola found transgenes in 80% of wild (uncultivated or "feral") varieties in North Dakota, meaning 80% of the plants which had established themselves in the area were genetically engineered varieties. The researchers stated that "we found the highest densities of [such transgene-containing] plants near agricultural fields and along major freeways, but we were also finding plants in the middle of nowhere" adding that "over time,..the build-up of different types of herbicide resistance in feral [natural] canola and closely related weeds, like field mustard, could make it more difficult to manage these plants using herbicides."

Chapter 3

Artificial Photosynthesis

Artificial photosynthesis is a research field that attempts to replicate the natural process of photosynthesis, converting sunlight, water, and carbon dioxide into carbohydrates and oxygen. Sometimes, splitting water into hydrogen and oxygen by using sunlight energy is also referred to as *artificial photosynthesis*. The actual process that allows half of the overall photosynthetic reaction to take place is photo-oxidation. This half-reaction is essential in separating water molecules because it releases hydrogen and oxygen ions. These ions are needed to reduce carbon dioxide into a fuel. However, the only known way this is possible is through an external catalyst, one that can react quickly as well as constantly absorb the sun's photons. The general basis behind this theory is the creation of an "artificial plant" type fuel source.

Research

The research being done can be split up into a series of approaches:

- The approach of the photoelectrochemical cell
- The approach of the dye-sensitized solar cell
- The approach of the NADP+/NADPH Coenzyme
- The approach of the Photocatalytic Water Splitting Under Solar Light

Photoelectrochemical cell

Research is being done into finding catalysts that can convert water, carbon dioxide, and sunlight to carbohydrates. For the first type of catalysts, nature usually uses the oxygen evolving complex. Having studied this complex, researchers have made catalysts such as blue dimer to mimic its function, but these catalysts were very inefficient. Another catalyst was engineered by Paul Kögerler, which uses four ruthenium atoms.

The carbohydrate-converting catalysts used in nature are the hydrogenases. Catalysts invented by engineers to mimic the hydrogenases include a catalyst by Cédric Tard, the rhodium atom catalyst from MIT, and the cobalt catalyst from MIT. Dr. Nocera of MIT is receiving funding from the Air Force Office of Scientific Research to help conduct the necessary experiments to push forward in catalyst research. The government funding has helped Nocera make the research possible and in turn he is providing them with strong results.

Advantages

- Dye-sensitized cells can be made at one-fifth of the price of silicium cells.
- The solar energy can be immediately converted and stored, unlike in PV cells, for example, which need to convert the energy and then store it into a battery (both operations implying energy losses). Furthermore, hydrogen as well as carbon-based storage options are quite environmentally friendly.
- Renewable, carbon-neutral source of energy, whether it is used for transportation or homes. Also the CO_2 emissions that have been distributed from fossil fuels will begin to diminish because of the photosynthetic properties of the reactions.

Disadvantages

- Artificial photosynthesis cells (currently) last no longer than a few years (unlike PV and passive solar panels, for example, which last twenty years or longer).
- The cost for alteration right now is not advantageous enough to compete with fossil fuels and natural gas as a viable source of mainstream energy.

Dye-sensitized solar cell

This form of cell uses a dye in the second part of the energy creation to separate the charges and create a current. Silicon is only used as a photoelectron source. The main advantages are lower production costs and a higher energy/cost ratio.

Research is also being done into a streamlined form of photosynthesis that breaks water into oxygen and hydrogen. This process is called photoelectrolysis. This process is the first stage of plant photosynthesis (the light-dependent reaction). Carbon dioxide is not consumed in this process. The hydrogen released by the eletrolysis could be used immediately to generate electricity or could be stored and used as a fuel.

The light-independent reaction (also known as the Calvin-Benson cycle) is the second stage of plant photosynthesis, which converts carbon dioxide into glucose. Glucose provides an energy store for the plant's growth and repair. It has been suggested that such a process replicated on an industrial scale could help offset CO_2 emissions and mitigate the risk of global warming. Specifically, the light-independent reaction of photosynthesis could be used to "mop up" excess carbon dioxide in the atmosphere. Photoelectrolysis or other renewable sources would have to provide the energy for this process in order for it to have net negative CO_2 emissions.

NADP+/NADPH Coenzyme

The coenzyme, behaving in a cyclic manner, goes between picking up a proton and two electrons. It then delivers the hydride to an area where they await the production of carbohydrates. The coenzyme in a natural photosynthetic cycle is recyclable, however the problem is this process cannot yet be replicated in a laboratory.

Right now the main aspiration for scientists is to obtain an NADPH-inspired catalyst capable of recreating the natural cyclic process. Utilizing light, hydride donors will be regenerated as well as produced where the molecules are continuously used. Brookhaven chemists are now using a ruthenium-based complex to serve as the acting model. The complex is proven to perform correspondingly with NADP+/NADPH, behaving as the foundation for the proton and two electrons needed to convert acetone to isopropanol.

Currently the Brookhaven researchers are aiming to find ways to have light generate the hydride donors. The general idea is to use this theory to produce fuels from carbon dioxide.

Photocatalytic Water Splitting Under Solar Light

Sustainable hydrogen production is a key target in the development of alternative energy systems of the future for providing a clean and affordable energy supply. The conversion of solar energy into hydrogen via a water-splitting process assisted by photosemiconductor catalysts is one of the most promising technologies for the future because large quantities of hydrogen can potentially be generated in a clean and sustainable manner. The conversion of solar energy into a clean fuel (H2) under ambient conditions is one of the greatest challenges facing scientists in the twenty-first century. This process is assisted by photocatalysts suspended directly in water instead of using photovoltaic and an electrolytic system, therefore the reaction is in just one step and it can be more efficient than Photoelectrochemical water splitting

Potential Global Impact

Artificial photosynthesis is a renewable, carbon-neutral source of fuel, producing either hydrogen, or carbohydrates. This sets it apart from the other popular renewable energy sources — hydroelectric, solar photovoltaic, geothermal, and wind — which produce electricity directly, with no fuel intermediate.

As such, artificial photosynthesis may become a very important source of fuel for transportation. Unlike biomass energy, it does not require arable land, and so it need not compete with the food supply.

Since the light-independent phase of photosynthesis fixes carbon dioxide from the atmosphere, artificial photosynthesis may provide an economical mechanism for carbon sequestration, reducing the pool of CO_2 in the atmosphere, and thus mitigating its effect on global warming. Specifically, net reduction of CO_2 will occur when artificial photosynthesis is used to produce carbon-based fuel which is stored indefinitely.

At the 15th meeting of the International Congress of Photosynthesis Research (ISPR) in Beijing 27 August 2010 a proposal was made for a macroscience Global Artificial Photosynthesis Project-seven models being presented for evaluation.

Time-line

1967

Akira Fujishima discovers the Honda-Fujishima effect in titanium dioxide, which can be used for hydrolysis.

2000

CSIRO press release on Artificial Photosynthesis

2003

Brookhaven National Laboratory press release

2006

SLAC on photogeneration cells

2008

MIT Chemist Daniel G. Nocera, head of the MIT's Solar Revolution Project, and postdoctoral fellow Matthew Kanan may have significantly reduced the cost of the materials required for splitting water into its constitute components by substituting expensive platinum for inexpensive cobalt and phosphate. This breakthrough may be combined with work being done by Chemist Bjorn Winther-Jensen of Monash University in Australia in developing a low cost conducting polymer that has a large surface area and is resistant to operational degradation. Such research may herald fuel cells that can perform useful work at lower energy thresholds over a longer lifecycle and at lower cost.

2009

Mexican, Japanese and Spanish publication show the impact of the thermal treatment in a closed atmosphere using Cd1-xZnxS photocatalysts. Cd1-xZnxS solid solution reports an incredible high activity in Hydrogen production from water splitting under sunlight irradiation.

2009

Leibniz Institute for Catalysis reports inexpensive iron carbonyl complexes

2009

University of East Anglia, A gold electrode covered with layers of indium phosphide nanoparticles for solar to hydrogen with 60% efficiency.

2010

Mitsubishi is developing its own artificial photosynthesis to use sunlight, water and carbon dioxide to "create the carbon building blocks from which resins, plastics and fibers can be synthesized."

2010

UCSC elemental doping and quantum dot sensitization show promise for solar hydrogen generation.
The Joint Center for Artificial Photosynthesis is set up by Caltech and Lawrence Berkeley National Laboratory and is funded by DOE as one of its Energy Innovation Hubs.

Chapter 4

Stem Cell Treatments

Stem cell treatments are a type of intervention strategy that introduces new cells into damaged tissue in order to treat disease or injury. Many medical researchers believe that stem cell treatments have the potential to change the face of human disease and alleviate suffering. The ability of stem cells to self-renew and give rise to subsequent generations with variable degrees of differentiation capacities, offers significant potential for generation of tissues that can potentially replace diseased and damaged areas in the body, with minimal risk of rejection and side effects.

A number of stem cell therapeutics exist, but most are at experimental stages and/or costly, with the notable exception of bone marrow transplantation. Medical researchers anticipate that adult and embryonic stem cells will soon be able to treat cancer, Type 1 diabetes mellitus, Parkinson's disease, Huntington's disease, Celiac Disease, cardiac failure, muscle damage and neurological disorders, and many others. Nevertheless, before stem cell therapeutics can be applied in the clinical setting, more research is necessary to understand stem cell behavior upon transplantation as well as the mechanisms of stem cell interaction with the diseased/injured microenvironment.

Current treatments

For over 30 years, bone marrow, and more recently, umbilical cord blood stem cells, have been used to treat cancer patients with conditions such as leukemia and lymphoma. During chemotherapy, most growing cells are killed by the cytotoxic agents. These agents, however, cannot discriminate between the leukemia or neoplastic cells, and the hematopoietic stem cells within the bone marrow. It is this side effect of conventional chemotherapy strategies that the stem cell transplant attempts to reverse; a donor's healthy bone marrow reintroduces functional stem cells to replace the cells lost in the host's body during treatment.

Potential treatments

Brain damage

Stroke and traumatic brain injury lead to cell death, characterized by a loss of neurons and oligodendrocytes within the brain. Healthy adult brains contain neural stem cells which divide to maintain general stem cell numbers, or become progenitor cells. In

healthy adult animals, progenitor cells migrate within the brain and function primarily to maintain neuron populations for olfaction (the sense of smell). Interestingly, in pregnancy and after injury, this system appears to be regulated by growth factors and can increase the rate at which new brain matter is formed. Although the reparative process appears to initiate following trauma to the brain, substantial recovery is rarely observed in adults, suggesting a lack of robustness.

Stem cells may also be used to treat brain degeneration, such as in Parkinson's and Alzheimer's disease.

Cancer

Research injecting neural (adult) stem cells into the brains of dogs has shown to be very successful in treating cancerous tumors. Using conventional techniques, brain cancer is difficult to treat because it spreads so rapidly. Researchers at the Harvard Medical School transplanted human neural stem cells into the brain of rodents that received intracranial tumours. Within days, the cells migrated into the cancerous area and produced cytosine deaminase, an enzyme that converts a non-toxic pro-drug into a chemotheraputic agent. As a result, the injected substance was able to reduce the tumor mass by 81 percent. The stem cells neither differentiated nor turned tumorigenic. Some researchers believe that the key to finding a cure for cancer is to inhibit proliferation of cancer stem cells. Accordingly, current cancer treatments are designed to kill cancer cells. However, conventional chemotherapy treatments cannot discriminate between cancerous cells and others. Stem cell therapies may serve as potential treatments for cancer. Research on treating Lymphoma using adult stem cells is underway and has had human trials. Essentially, chemotherapy is used to completely destroy the patients own lymphocytes, and stem cells injected, eventually replacing the immune system of the patient with that of the healthy donor.

Spinal cord injury

A team of Korean researchers reported on November 25, 2003, that they had transplanted multipotent adult stem cells from umbilical cord blood to a patient suffering from a spinal cord injury and that following the procedure, she could walk on her own, without difficulty. The patient had not been able to stand up for roughly 19 years. For the unprecedented clinical test, the scientists isolated adult stem cells from umbilical cord blood and then injected them into the damaged part of the spinal cord.

According to the October 7, 2005 issue of *The Week*, University of California, Irvine researchers transplanted multipotent human fetal-derived neural stem cells into paralyzed mice, resulting in locomotor improvements four months later. The observed recovery was associated with differentiation of transplanted cells into new neurons and oligodendrocytes- the latter of which forms the myelin sheath around axons of the central nervous system, thus insulating neural impulses and facilitating communication with the brain.

In January 2005, researchers at the University of Wisconsin–Madison differentiated human blastocyst stem cells into neural stem cells, then into pre-mature motor neurons, and finally into spinal motor neurons, the cell type that, in the human body, transmits messages from the brain to the spinal cord and subsequently mediates motor function in the periphery. The newly generated motor neurons exhibited electrical activity, the signature action of neurons. Lead researcher Su-Chun Zhang described the process as "teaching the blastocyst stem cells to change step by step, where each step has different conditions and a strict window of time."

Transformation of blastocyst stem cells into motor neurons had eluded researchers for decades. While Zhang's findings were a significant contribution to the field, the ability of transplanted neural cells to establish communication with neighboring cells remains unclear. Accordingly, studies using chicken embryos as a model organism can be an effective proof-of-concept experiment. If functional, the new cells could be used to treat diseases like Lou Gehrig's disease, muscular dystrophy, and spinal cord injuries.

Heart damage

Several clinical trials targeting heart disease have shown that adult stem cell therapy is safe, effective, and equally efficient in treating old and recent infarcts. Adult stem cell therapy for treating heart disease was commercially available in at least five continents at the last count (2007).

Possible mechanisms of recovery include:

- Generation of heart muscle cells
- Stimulation of growth of new blood vessels to repopulate damaged heart tissue
- Secretion of growth factors
- Assistance via some other mechanism

It may be possible to have adult bone marrow cells differentiate into heart muscle cells.

Haematopoiesis (blood cell formation)

The specificity of the human immune cell repertoire is what allows the human body to defend itself from rapidly adapting antigens. However, the immune system is vulnerable to degradation upon the pathogenesis of disease, and because of the critical role that it plays in overall defense, its degradation is often fatal to the organism as a whole. Diseases of hematopoietic cells are called hematopathology. The specificity of the immune cells is what allows recognition of foreign antigens, causing further challenges in the treatment of immune disease. Identical matches between donor and recipient must be made for successful transplantation treatments, but matches are uncommon, even between first-degree relatives. Research using both hematopoietic adult stem cells and embryonic stem cells has provided insight into the possible mechanisms and methods of treatment for many of these ailments.

Fully mature human red blood cells may be generated *ex vivo* by hematopoietic stem cells (HSCs), which are precursors of red blood cells. In this process, HSCs are grown together with stromal cells, creating an environment that mimics the conditions of bone marrow, the natural site of red blood cell growth. Erythropoietin, a growth factor, is added, coaxing the stem cells to complete terminal differentiation into red blood cells. Further research into this technique should have potential benefits to gene therapy, blood transfusion, and topical medicine.

Baldness

Hair follicles also contain stem cells, and some researchers predict research on these follicle stem cells may lead to successes in treating baldness through an activation of the stem cells progenitor cells. This treatment is expected to work by activating already existing stem cells on the scalp. Later treatments may be able to simply signal follicle stem cells to give off chemical signals to nearby follicle cells which have shrunk during the aging process, which in turn respond to these signals by regenerating and once again making healthy hair.

Missing teeth

In 2004, scientists at King's College London discovered a way to cultivate a complete tooth in mice and were able to grow them stand-alone in the laboratory. Researchers are confident that this technology can be used to grow live teeth in human patients.

In theory, stem cells taken from the patient could be coaxed in the lab into turning into a tooth bud which, when implanted in the gums, will give rise to a new tooth, and would be expected to grow within two months. It will fuse with the jawbone and release chemicals that encourage nerves and blood vessels to connect with it. The process is similar to what happens when humans grow their original adult teeth.

Many challenges remain, however, before stem cells could be a choice for the replacement of missing teeth in the future.

Deafness

Heller has reported success in re-growing cochlea hair cells with the use of embryonic stem cells.

Blindness and vision impairment

Since 2003, researchers have successfully transplanted corneal stem cells into damaged eyes to restore vision. Using embryonic stem cells, scientists are able to grow a thin sheet of totipotent stem cells in the laboratory. When these sheets are transplanted over the damaged cornea, the stem cells stimulate renewed repair, eventually restore vision. The latest such development was in June 2005, when researchers at the Queen Victoria Hospital of Sussex, England were able to restore the sight of forty patients using the same

technique. The group, led by Dr. Sheraz Daya, was able to successfully use adult stem cells obtained from the patient, a relative, or even a cadaver. Further rounds of trials are ongoing.

In April 2005, doctors in the UK transplanted corneal stem cells from an organ donor to the cornea of Deborah Catlyn, a woman who was blinded in one eye when acid was thrown in her eye at a nightclub. The cornea, which is the transparent window of the eye, is a particularly suitable site for transplants. In fact, the first successful human transplant was a cornea transplant. The absence of blood vessels within the cornea makes this area a relatively easy target for transplantation. The majority of corneal transplants carried out today are due to a degenerative disease called keratoconus.

The University Hospital of New Jersey reports that the success rate for growth of new cells from transplanted stem cells varies from 25 percent to 70 percent.

In 2009, researchers at the University of Pittsburgh Medical center demonstrated that stem cells collected from human corneas can restore transparency without provoking a rejection response in mice with corneal damage.

Amyotrophic lateral sclerosis

Stem cells have resulted in significant locomotor improvements in rats with an Amyotrophic lateral sclerosis-like disease. In a rodent model that closely mimics the human form of ALS, animals were injected with a virus to kill the spinal cord motor nerves which mediate movement. Animals subsequently received stem cells in the spinal cord. Transplanted cells migrated to the sites of injury, contributed to regeneration of the ablated nerve cells, and restored locomotor function.

Graft vs. host disease and Crohn's disease

Phase III clinical trials expected to end in second-quarter 2008 were conducted by Osiris Therapeutics using their in-development product Prochymal, derived from adult bone marrow. The target disorders of this therapeutic are graft-versus-host disease and Crohn's disease.

Neural and behavioral birth defects

A team of researchers led by Prof. Joseph Yanai were able to reverse learning deficits in the offspring of pregnant mice who were exposed to heroin and the pesticide organophosphate. This was done by direct neural stem cell transplantation into the brains of the offspring. The recovery was almost 100 percent, as shown in behavioral tests that suggested improved to normal behavior and learning scores in animals receiving cell transplantation. On the molecular level, brain chemistry of the treated animals was also restored to normal. Through the work, which was supported by the US National Institutes of Health, the US-Israel Binational Science Foundation and the Israel anti-drug

authorities, the researchers discovered that the stem cells worked even in cases where most of the cells died out in the host brain.

The scientists found that before they die the neural stem cells succeed in inducing the host brain to produce large numbers of stem cells which repair the damage. These findings, which answered a major question in the stem cell research community, were published earlier this year in the leading journal, Molecular Psychiatry. Scientists are now developing procedures to administer the neural stem cells in the least invasive way possible - probably via blood vessels, making therapy practical and clinically feasible. Researchers also plan to work on developing methods to take cells from the patient's own body, turn them into stem cells, and then transplant them back into the patient's blood via the blood stream. Aside from decreasing the chances of immunological rejection, the approach will also eliminate the controversial ethical issues involved in the use of stem cells from human embryos.

Diabetes

Diabetes patients lose the function of insulin-producing beta cells within the pancreas. Human embryonic stem cells may be grown in cell culture and stimulated to form insulin-producing cells that can be transplanted into the patient.

However, clinical success is highly dependent on the development of the following procedures:

- Transplanted cells should proliferate
- Transplanted cells should differentiate in a site-specific manner
- Transplanted cells should survive in the recipient (prevention of transplant rejection)
- Transplanted cells should integrate within the targeted tissue
- Transplanted cells should integrate into the host circuitry and restore function

Orthopaedics

Clinical case reports in the treatment of orthopaedic conditions have been reported. To date, the focus in the literature for musculoskeletal care appears to be on mesenchymal stem cells. Centeno et al. have published MRI evidence of increased cartilage and meniscus volume in individual human subjects. The results of trials that include a large number of subjects, are yet to be published. However, a published safety study conducted in a group of 227 patients over a 3-4 year period shows adequate safety and minimal complications associated with mesenchymal cell transplantation.

Wakitani has also published a small case series of nine defects in five knees involving surgical transplantation of mesenchymal stem cells with coverage of the treated chondral defects.

Wound healing

Stem cells can also be used to stimulate the growth of human tissues. In an adult, wounded tissue is most often replaced by scar tissue, which is characterized in the skin by disorganized collagen structure, loss of hair follicles and irregular vascular structure. In the case of wounded fetal tissue, however, wounded tissue is replaced with normal tissue through the activity of stem cells. A possible method for tissue regeneration in adults is to place adult stem cell "seeds" inside a tissue bed "soil" in a wound bed and allow the stem cells to stimulate differentiation in the tissue bed cells. This method elicits a regenerative response more similar to fetal wound-healing than adult scar tissue formation. Researchers are still investigating different aspects of the "soil" tissue that are conducive to regeneration.

Infertility

Culture of human embryonic stem cells in mitotically inactivated porcine ovarian fibroblasts (POF) causes differentiation into germ cells (precursor cells of oocytes and spermatozoa), as evidenced by gene expression analysis.

Human embryonic stem cells have been stimulated to form Spermatozoon-like cells, yet still slightly damaged or malformed. It could potentially treat azoospermia.

Clinical Trials

On January 23, 2009, the US Food and Drug Administration gave clearance to Geron Corporation for the initiation of the first clinical trial of an embryonic stem cell-based therapy on humans. The trial will evaluate the drug GRNOPC1, embryonic stem cell-derived oligodendrocyte progenitor cells, on patients with acute spinal cord injury.

As of mid 2010 hundreds of phase III clinical trials involving stem cells have been registered.

Stem cell use in animals

Veterinary applications

Potential contributions to veterinary medicine

Research currently conducted on horses, dogs, and cats can benefit the development of stem-cell treatments in veterinary medicine and can target a wide range of injuries and diseases such as myocardial infarction, stroke, tendon and ligament damage, osteoarthritis, osteochondrosis and muscular dystrophy both in large animals, as well as humans. While investigation of cell-based therapeutics generally reflects human medical needs, the high degree of frequency and severity of certain injuries in racehorses has put veterinary medicine at the forefront of this novel regenerative approach. Companion animals can serve as clinically relevant models that closely mimic human disease.

Development of regenerative treatment models

Veterinary applications of stem cell therapy as a means of tissue regeneration have been largely shaped by research that began with the use of adult-derived mesenchymal stem cells to treat animals with injuries or defects affecting bone, cartilage, ligaments and/or tendons. Because mesenchymal stem cells can differentiate into the cells that make up bone, cartilage, tendons, and ligaments (as well as muscle, fat, and possibly other tissues), they have been the main type of stem cells studied in the treatment of diseases affecting these tissues. Mesenchymal stem cells are primarily derived from adipose tissue or bone marrow. Since an elevated immune response following cell transplantation may result in rejection of exogenous cells (except in the case of cells derived from a very closely genetically related individual), mesenchymal stem cells are often derived from the patient prior to injection in a process known as autologous transplantation. Surgical repair of bone fractures in dogs and sheep has demonstrated that engraftment of mesenchymal stem cells derived from a genetically different donor within the same species, termed allogeneic transplantation, does not elicit an immunological response in the recipient animal and can mediate regeneration of bone tissue in major bony fractures and defects. Stem cells can speed up bone repair in fractures/defects that would normally require extensive grafting, suggesting that mesenchymal stem cell use may provide a useful alternative to conventional grafting techniques. Treating tendon and ligament injuries in horses using stem cells, whether derived from adipose tissue or bone-marrow, has support in the veterinary literature. While further studies are necessary to fully characterize the use of cell-based therapeutics for treatment of bone fractures, stem cells are thought to mediate repair via five primary mechanisms: 1) providing an antiinflammatory effect, 2) homing to damaged tissues and recruiting other cells, such as endothelial progenitor cells, that are necessary for tissue growth, 3) supporting tissue remodeling over scar formation, 4) inhibiting apoptosis, and 5) differentiating into bone, cartilage, tendon, and ligament tissue.

Significance of stem cell microenvironments

The microenvironment into which stem cells are transplanted significantly alters the capacity of engrafted cells for recovery and repair. The microenviroment provides growth factors and other chemical signals that guide appropriate differentiation of transplanted cell populations and direct transplanted cells to sites of trauma or disease. Repair and recovery can then be mediated via three primary mechanisms: 1) formation and/or recruitment of new blood cells to the damaged region; 2) prevention of programed cell death or apoptosis; and 3) suppression of inflammation. To further enrich blood supply to the damaged areas, and consequently promote tissue regeneration, platelet-rich plasma could be used in conjunction with stem cell transplantation. The efficacy of some stem cell populations may also be affected by the method of delivery; for instance, to regenerate bone, stem cells are often introduced in a scaffold where they produce the minerals necessary for generation of functional bone.

Sources of autologous (patient-derived) stem cells

Autologous stem cells intended for regenerative therapy are generally isolated either from the patient's bone marrow or from adipose tissue. The number of stem cells transplanted into damaged tissue may alter efficacy of treatment. Accordingly, stem cells derived from bone marrow aspirates, for instance, are cultured in specialized laboratories for expansion to millions of cells. Although adipose-derived tissue also requires processing prior to use, the culturing methodology for adipose-derived stem cells is not as extensive as that for bone marrow-derived cells. While it is thought that bone-marrow derived stem cells are preferred for bone, cartilage, ligament, and tendon repair, others believe that the less challenging collection techniques and the multi-cellular microenvironment already present in adipose-derived stem cell fractions make the latter the preferred source for autologous transplantation.

Currently Available Treatments for Horses and Dogs Suffering from Orthopedic Conditions

Autologous or allogeneic stem cells are currently used as an adjunctive therapy in the surgical repair of some types of fractures in dogs and horses. Autologous stem cell-based treatments for ligament injury, tendon injury, osteoarthritis, osteochondrosis, and sub-chondral bone cysts have been commercially available to practicing veterinarians to treat horses since 2003 in the United States and since 2006 in the United Kingdom. Autologous stem-cell based treatments for tendon injury, ligament injury, and osteoarthritis in dogs have been available to veterinarians in the United States since 2005. Over 3000 privately-owned horses and dogs have been treated with autologous adipose-derived stem cells. The efficacy of these treatments has been shown in double-blind clinical trials for dogs with osteoarthritis of the hip and elbow and horses with tendon damage. The efficacy of using stem cells, whether adipose-derived or bone-marrow derived, for treating tendon and ligament injuries in horses has support in the veterinary literature.

Developments in Stem Cell Treatments in Veterinary Internal Medicine

Currently, research is being conducted to develop stem cell treatments for: 1) horses suffering from COPD, neurologic disease, and laminitis; and 2) dogs and cats suffering from heart disease, liver disease, kidney disease, neurologic disease, and immune-mediated disorders.

Embryonic stem cell controversy

There is widespread controversy over the use of human embryonic stem cells. This controversy primarily targets the techniques used to derive new embryonic stem cell lines, which often requires the destruction of the blastocyst.

Opposition to the use of human embryonic stem cells in research is often based on philosophical, moral or religious objections. While these objections are the source of

rhetoric opposing the research, those opposed to such research can also point to the complete failure of embryonic stem cell researchers to provide anything of substance.

There have been tens of thousands of successful adult stem cell treatments just in China over the last decade, while the research community has yet to produce even one positive patient result using embryonic stem cells. The alternatives do not require the destruction of an embryo, such as the use of umbilical cord blood, milk teeth stem cells, bone marrow stem cells or using induced pluripotent stem cells. The stem cells from these alternative sources also lack the severe side effects which are universally seen in embryonic stem cell therapies and most often result in fatal rejection by the subject; the most common being mutations of the stem cells into tumors.

Stem cell treatments around the world

China

Stem cell research and treatment is currently being practiced at a clinical level in the People's Republic of China. The Ministry of Health of the People's Republic of China has permitted the use of stem cell therapy for conditions beyond those approved of in Western countries such as the United States, United Kingdom, and Australia. The Western World has scrutinized China for its failed attempts to meet international documentation standards of these trials and procedures, despite the overwhelmingly positive anecdotal results.

Stem cell therapies provided in China utilize a variety of cell types including umbilical cord stem cells and olfactory ensheathing cells. The stem cells are then expanded in centralized blood banks before being used in stem cell treatments. State-funded companies based in the Shenzhen Hi-Tech Industrial Zone treat the symptoms of numerous disorders with adult stem cell therapy. Hospitals throughout eastern China provide numerous therapies to patients in coordination with the stem cell providers. These companies' therapies are currently focused on the treatment of neurodegenerative and cardiovascular disorders. The most radical successes of Chinese adult stem cell therapy have been in treating the brain. These therapies administer stem cells directly to the brain to promote greater motor and brain function in patients with Cerebral Palsy, Alzheimer's, and brain injuries. However, retrospective studies have shown that Chinese use of fetal-derived brain tissue in spinal cord injured human subjects were not as promising as once thought: the phenotype and the fate of the transplanted cells, described as olfactory ensheathing cells, were unknown. As well, perioperative morbidity and lack of functional benefit were identified as the most serious clinical shortcomings. Furthermore, the extent of regulatory policy in the use of stem cell therapies in China is unclear. In the absence of a valid clinical trials protocol, and more regulatory oversight, Western regulatory agencies advise patients and physicians to be cautious when selecting Chinese stem cell therapeutic centers.

Mexico

Stem cell treatment is currently being practiced at a clinical level in Mexico. An International Health Department Permit (COFEPRIS) is required. This permit allows the use of stem cell types beyond those approved of in Western countries such as the United States or Europe. Stem cell therapies provided in Mexico utilize patient Adipose, Bone Marrow, or Donor Placenta sources.

South Korea

In 2005, South Korean scientists claimed to have generated stem cells that were tailored to match the recipient. Each of the 11 new stem cell lines was developed using somatic cell nuclear transfer (SCNT) technology. The resultant cells were thought to match the genetic material of the recipient, thus suggesting minimal to no cell rejection.

This study, however, was eventually discredited as the primary researcher,Dr. Woo Suk Hwang, admitted to using cells obtained from his research staff. In Dec 2005, claims were put forward that his research had been manipulated to wrongfully indicate positive results. Later that month, these claims were confirmed by an academic panel.

Chapter 5

Hydrogen Economy

The **hydrogen economy** is a proposed system of delivering energy using hydrogen. The term *hydrogen economy* was coined by John Bockris during a talk he gave in 1970 at General Motors (GM) Technical Center.

Hydrogen advocates promote hydrogen as potential fuel for motive power (including cars and boats), the energy needs of buildings and portable electronics. Free hydrogen does not occur naturally in quantity, and thus it must be generated from some other energy source by steam reformation of natural gas or another method. Hydrogen is therefore an energy carrier (like electricity), not a primary energy source (like coal). The utility of a hydrogen economy depends on issues of energy sourcing, including fossil fuel use, climate change, and sustainable energy generation.

Rationale

Elements of the hydrogen economy

A hydrogen economy is proposed to solve some of the negative effects of using hydrocarbon fuels where the carbon is released to the atmosphere. Modern interest in the hydrogen economy can generally be traced to a 1970 technical report by Lawrence W. Jones of the University of Michigan.

In the current hydrocarbon economy, transportation is fueled primarily by petroleum. Burning of hydrocarbon fuels emits carbon dioxide and other pollutants. The supply of economically usable hydrocarbon resources in the world is limited, and the demand for hydrocarbon fuels is increasing, particularly in China, India and other developing countries.

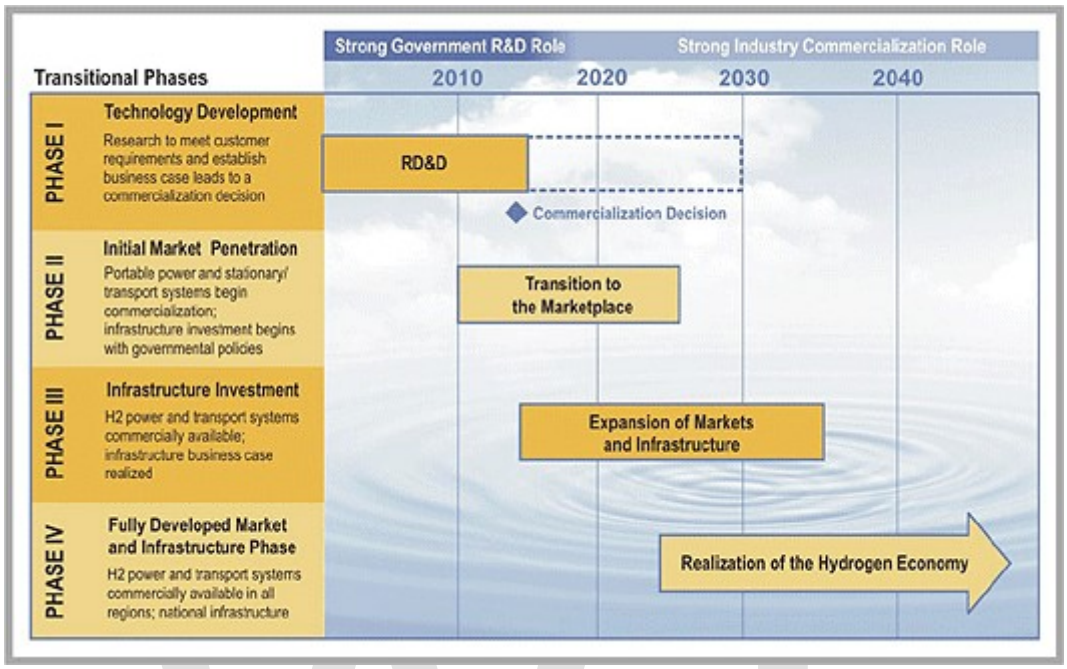
Proponents of a world-scale hydrogen economy argue that hydrogen can be an environmentally cleaner source of energy to end-users, particularly in transportation applications, without release of pollutants (such as particulate matter) or carbon dioxide at the point of end use. A 2004 analysis asserted that "most of the hydrogen supply chain pathways would release significantly less carbon dioxide into the atmosphere than would gasoline used in hybrid electric vehicles" and that significant reductions in carbon dioxide emissions would be possible if carbon capture or carbon sequestration methods were utilized at the site of energy or hydrogen production.

Hydrogen has a high energy density by weight. An Otto cycle internal combustion engine running on hydrogen is said to have a maximum efficiency of about 38%, 8% higher than gasoline internal combustion engine.

The combination of the fuel cell and electric motor is 2-3 times more efficient than an internal combustion engine. However, the high capital costs of fuel cells, about $5,500/kW in 2002, are one of the major obstacles of its development, meaning that the fuel cell is only technically, but not economically, more efficient than an internal combustion engine.

Other technical obstacles include hydrogen storage issues and the purity requirement of hydrogen used in fuel cells – with current technology, an operating fuel cell requires the purity of hydrogen to be as high as 99.999%. On the other hand, hydrogen engine conversion technology is more economical than fuel cells.

Perspective: current hydrogen market (current hydrogen economy)



Timeline

Hydrogen production is a large and growing industry. Globally, some 50 million metric tons of hydrogen, equal to about 170 million tons of oil equivalent, were produced in 2004. The growth rate is around 10% per year. Within the United States, 2004 production was about 11 million metric tons (MMT), an average power flow of 48 gigawatts. (For comparison, the average electric production in 2003 was some 442 gigawatts.) As of 2005, the economic value of all hydrogen produced worldwide is about $135 billion per year.

There are two primary uses for hydrogen today. About half is used to produce ammonia (NH_3) via the Haber process, which is then used directly or indirectly as fertilizer. Because both the world population and the intensive agriculture used to support it are growing, ammonia demand is growing. The other half of current hydrogen production is used to convert heavy petroleum sources into lighter fractions suitable for use as fuels. This latter process is known as hydrocracking. Hydrocracking represents an even larger growth area, since rising oil prices encourage oil companies to extract poorer source material, such as tar sands and oil shale. The scale economies inherent in large scale oil refining and fertilizer manufacture make possible on-site production and "captive" use. Smaller quantities of "merchant" hydrogen are manufactured and delivered to end users as well.

If energy for hydrogen production were available (from wind, solar or nuclear power), use of the substance for hydrocarbon synfuel production could expand captive use of hydrogen by a factor of 5 to 10. Present U.S. use of hydrogen for hydrocracking is

roughly 4 million metric tons per year (4 MMT/yr). It is estimated that 37.7 MMT/yr of hydrogen would be sufficient to convert enough domestic coal to liquid fuels to end U.S. dependence on foreign oil importation, and less than half this figure to end dependence on Middle East oil. Coal liquefaction would present significantly worse emissions of carbon dioxide than does the current system of burning fossil petroleum, but it would eliminate the political and economic vulnerabilities inherent in oil importation.

Currently, global hydrogen production is 48% from natural gas, 30% from oil, and 18% from coal; water electrolysis accounts for only 4%. The distribution of production reflects the effects of thermodynamic constraints on economic choices: of the four methods for obtaining hydrogen, partial combustion of natural gas in a NGCC (natural gas combined cycle) power plant offers the most efficient chemical pathway and the greatest off-take of usable heat energy.

The large market and sharply rising prices in fossil fuels have also stimulated great interest in alternate, cheaper means of hydrogen production. As of 2002, most hydrogen is produced on site and the cost is approximately $0.32/lb and, if not produced on site, the cost of liquid hydrogen is about $1.00/lb to $1.40/lb.

Production, storage, infrastructure

Today hydrogen is mainly produced (90%) from fossil sources. Linking its centralized production to a fleet of light-duty fuel cell vehicles will require the siting and construction of a distribution infrastructure with large investment of capital. Further, the technological challenge of providing safe, energy-dense storage of hydrogen on-board the vehicle must be overcome to provide sufficient range between fillups.

Methods of production

Molecular hydrogen is not available on Earth in convenient natural reservoirs. Most hydrogen on Earth is bonded to oxygen in water. Manufacturing elemental hydrogen does require the consumption of a hydrogen carrier such as a fossil fuel or water. The former consumes the fossil resource and produces carbon dioxide, but often requires no further energy input beyond the fossil fuel. Decomposing water requires electrical or heat input, generated from some primary energy source (fossil fuel, nuclear power or a renewable energy).

Current production methods

Hydrogen is industrially produced from steam reforming, which uses fossil fuels such as natural gas, oil, or coal. The energy content of the produced hydrogen is less than the energy content of the original fuel, some of it being lost as excessive heat during production. Steam reforming leads to carbon dioxide emissions, in the same way as a car engine would do.

A small part (4% in 2006) is produced by electrolysis using electricity and water, consuming approximately 50 kilowatt-hours of electricity per kilogram of hydrogen produced.

Kværner-process

The Kværner-process or Kvaerner carbon black & hydrogen process (CB&H) is a method, developed in the 1980s by a Norwegian company of the same name, for the production of hydrogen from hydrocarbons (C_nH_m), such as methane, natural gas and biogas. Of the available energy of the feed, approximately 48% is contained in the Hydrogen, 40% is contained in activated carbon and 10% in superheated steam.

Biological production

Fermentative hydrogen production is the fermentative conversion of organic substrate to biohydrogen manifested by a diverse group bacteria using multi enzyme systems involving three steps similar to anaerobic conversion. Dark fermentation reactions do not require light energy, so they are capable of constantly producing hydrogen from organic compounds throughout the day and night. Photofermentation differs from dark fermentation because it only proceeds in the presence of light. For example photo-fermentation with Rhodobacter sphaeroides SH2C can be employed to convert small molecular fatty acids into hydrogen. Electrohydrogenesis is used in microbial fuel cells where hydrogen is produced from organic matter (e.g. from sewage, or solid matter) while 0.2 - 0.8 V is applied.

Biological hydrogen can be produced in an algae bioreactor. In the late 1990s it was discovered that if the algae is deprived of sulfur it will switch from the production of oxygen, i.e. normal photosynthesis, to the production of hydrogen.

Biological hydrogen can be produced in bioreactors that use feedstocks other than algae, the most common feedstock being waste streams. The process involves bacteria feeding on hydrocarbons and excreting hydrogen and CO_2. The CO_2 can be sequestered successfully by several methods, leaving hydrogen gas. A prototype hydrogen bioreactor using waste as a feedstock is in operation at Welch's grape juice factory in North East, Pennsylvania.

Biocatalysed electrolysis

Besides regular electrolysis, electrolysis using microbes is another possibility. With biocatalysed electrolysis, hydrogen is generated after running through the microbial fuel cell and a variety of aquatic plants can be used. These include reed sweetgrass, cordgrass, rice, tomatoes, lupines, algae

Electrolysis of water

Electrolysis of water ship Hydrogen Challenger

Hydrogen can be made via high pressure electrolysis or low pressure electrolysis of water. Current best processes have an efficiency of 50% to 80%, so that 1 kg of hydrogen requires 50 to 80 kWh of electricity. At 8 cents/kWh, that's $4.00/kg, which is 3 to 10 times the price of hydrogen from steam reformation of natural gas. The price difference is due to the efficiency of direct conversion of fossil fuels to produce hydrogen, rather than burning fuel to produce electricity. Hydrogen from natural gas, used to replace e.g. gasoline, emits more CO2 than the gasoline it would replace, and so is no help in reducing greenhouse gases.

High-pressure electrolysis

High pressure electrolysis is the electrolysis of water by decomposition of water (H_2O) into oxygen (O_2) and hydrogen gas (H_2) by means of an electric current being passed through the water. The difference with a standard electrolyzer is the compressed hydrogen output around 120-200 Bar (1740-2900 psi). By pressurising the hydrogen in the electrolyser the need for an external hydrogen compressor is eliminated, the average energy consumption for internal compression is around 3%.

High-temperature electrolysis

Hydrogen can be generated from energy supplied in the form of heat and electricity through high-temperature electrolysis (HTE). Because some of the energy in HTE is supplied in the form of heat, less of the energy must be converted twice (from heat to electricity, and then to chemical form), and so potentially far less energy is required per kilogram of hydrogen produced.

While nuclear-generated electricity could be used for electrolysis, nuclear heat can be directly applied to split hydrogen from water. High temperature (950–1000 °C) gas cooled nuclear reactors have the potential to split hydrogen from water by thermochemical means using nuclear heat. Research into high-temperature nuclear reactors may eventually lead to a hydrogen supply that is cost-competitive with natural gas steam reforming. General Atomics predicts that hydrogen produced in a High Temperature Gas Cooled Reactor (HTGR) would cost $1.53/kg. In 2003, steam reforming of natural gas yielded hydrogen at $1.40/kg. At 2005 natural gas prices, hydrogen costs $2.70/kg.

High-temperature electrolysis has been demonstrated in a laboratory, at 108 megajoules (thermal) per kilogram of hydrogen produced, but not at a commercial scale. In addition, this is lower-quality "commercial" grade Hydrogen, unsuitable for use in fuel cells.

Photoelectrochemical water splitting

Using electricity produced by photovoltaic systems offers the cleanest way to produce hydrogen. Water is broken into hydrogen and oxygen by electrolysis—a photoelectrochemical cell (PEC) process which is also named artificial photosynthesis. Research aimed toward developing higher-efficiency multijunction cell technology is underway by the photovoltaic industry.

Concentrating solar thermal

Very high temperatures are required to dissociate water into hydrogen and oxygen. A catalyst is required to make the process operate at feasible temperatures. Heating the water can be achieved through the use of concentrating solar power. Hydrosol-2 is a 100-kilowatt pilot plant at the Plataforma Solar de Almería in Spain which uses sunlight to obtain the required 800 to 1,200 °C to heat water. Hydrosol II has been in operation since 2008. The design of this 100-kilowatt pilot plant is based on a modular concept. As a result, it may be possible that this technology could be readily scaled up to the megawatt range by multiplying the available reactor units and by connecting the plant to heliostat fields (fields of sun-tracking mirrors) of a suitable size.

Photoelectrocatalytic production

A method studied by Thomas Nann and his team at the University of East Anglia consists of a gold electrode covered in layers of indium phosphide (InP) nanoparticles. They introduced an iron-sulfur complex into the layered arrangement, which when submerged in water and irradiated with light under a small electric current, produced hydrogen with an efficiency of 60%.

Thermochemical production

There are more than 352 thermochemical cycles which can be used for water splitting, around a dozen of these cycles such as the iron oxide cycle, cerium(IV) oxide-cerium(III)

oxide cycle, zinc zinc-oxide cycle, sulfur-iodine cycle, copper-chlorine cycle and hybrid sulfur cycle are under research and in testing phase to produce hydrogen and oxygen from water and heat without using electricity. These processes can be more efficient than high-temperature electrolysis, typical in the range from 35 % - 49 % LHV efficiency. Thermochemical production of hydrogen using chemical energy from coal or natural gas is generally not considered, because the direct chemical path is more efficient.

None of the thermochemical hydrogen production processes have been demonstrated at production levels, although several have been demonstrated in laboratories.

Storage

Although molecular hydrogen has very high energy density on a mass basis, partly because of its low molecular weight, as a gas at ambient conditions it has very low energy density by volume. If it is to be used as fuel stored on board the vehicle, pure hydrogen gas must be pressurized or liquefied to provide sufficient driving range. Increasing gas pressure improves the energy density by volume, making for smaller, but not lighter container tanks. Achieving higher pressures necessitates greater use of external energy to power the compression. Alternatively, higher volumetric energy density liquid hydrogen or slush hydrogen may be used. However, liquid hydrogen is cryogenic and boils at 20.268 K (–252.882 °C or –423.188 °F). Cryogenic storage cuts weight but requires large liquification energies. The liquefaction process, involving pressurizing and cooling steps, is energy intensive. The liquefied hydrogen has lower energy density by volume than gasoline by approximately a factor of four, because of the low density of liquid hydrogen — there is actually more hydrogen in a liter of gasoline (116 grams) than there is in a liter of pure liquid hydrogen (71 grams). Liquid hydrogen storage tanks must also be well insulated to minimize boil off. Ice may form around the tank and help corrode it further if the liquid hydrogen tank insulation fails.

The mass of the tanks needed for compressed hydrogen reduces the fuel economy of the vehicle. Because it is a small molecule, hydrogen tends to diffuse through any liner material intended to contain it, leading to the embrittlement, or weakening, of its container.

Distinct from storing molecular hydrogen, hydrogen can be stored as a chemical hydride or in some other hydrogen-containing compound. Hydrogen gas is reacted with some other materials to produce the hydrogen storage material, which can be transported relatively easily. At the point of use the hydrogen storage material can be made to decompose, yielding hydrogen gas. As well as the mass and volume density problems associated with molecular hydrogen storage, current barriers to practical storage schemes stem from the high pressure and temperature conditions needed for hydride formation and hydrogen release. For many potential systems hydriding and dehydriding kinetics and heat management are also issues that need to be overcome.

A third approach is to absorb molecular hydrogen into a solid storage material. Unlike in the hydrides mentioned above, the hydrogen does not dissociate/recombine upon

charging/discharging the storage system, and hence does not suffer from the kinetic limitations of many hydride storage systems. Hydrogen densities similar to liquefied hydrogen can be achieved with appropriate absorption media. Some suggested absorbers include MOFs, nanostructured carbons (including CNTs) and clathrate hydrate.

The most common method of on board hydrogen storage in today's demonstration vehicles is as a compressed gas at pressures of roughly 700 bar (70 MPa).

Underground hydrogen storage is the practice of hydrogen storage in underground caverns, salt domes and depleted oil and gas fields. Large quantities of gaseous hydrogen are stored in underground caverns by ICI for many years without any difficulties. The storage of large quantities of hydrogen underground can function as grid energy storage which is essential for the hydrogen economy.

Infrastructure

Praxair Hydrogen Plant

The hydrogen infrastructure consists mainly of industrial hydrogen pipeline transport and hydrogen-equipped filling stations like those found on a hydrogen highway. Hydrogen stations which are not situated near a hydrogen pipeline get supply via hydrogen tanks, compressed hydrogen tube trailers, liquid hydrogen trailers, liquid hydrogen tank trucks or dedicated onsite production.

Because of hydrogen embrittlement of steel, and corrosion natural gas pipes require internal coatings or replacement in order to convey hydrogen. Techniques are well-known; over 700 miles of hydrogen pipeline currently exist in the United States. Although expensive, pipelines are the cheapest way to move hydrogen. Hydrogen gas piping is routine in large oil-refineries, because hydrogen is used to hydrocrack fuels from crude oil.

Hydrogen piping can in theory be avoided in distributed systems of hydrogen production, where hydrogen is routinely made on site using medium or small-sized generators which would produce enough hydrogen for personal use or perhaps a neighborhood. In the end, a combination of options for hydrogen gas distribution may succeed.

While millions of tons of elemental hydrogen are distributed around the world each year in various ways, bringing hydrogen to individual consumers would require an evolution of the fuel infrastructure. For example, according to GM, 70% of the U.S. population lives near a hydrogen-generating facility but has little public access to that hydrogen. The same study however, shows that building the infrastructure in a systematic way is much more doable and affordable than most people think. For example, one article has noted that hydrogen stations could be put within every 10 miles in metro Los Angeles, and on the highways between LA and neighboring cities like Palm Springs, Las Vegas, San Diego and Santa Barbara, for the cost of a Starbuck's latte for every one of the 15 million residents living in these areas.

A key tradeoff: centralized vs. distributed production

In a future full hydrogen economy, primary energy sources and feedstock would be used to produce hydrogen gas as stored energy for use in various sectors of the economy. Producing hydrogen from primary energy sources other than coal, oil, and natural gas, would result in lower production of the greenhouse gases characteristic of the combustion of these fossil energy resources.

One key feature of a hydrogen economy is that in mobile applications (primarily vehicular transport) energy generation and use is decoupled. The primary energy source need no longer travel with the vehicle, as it currently does with hydrocarbon fuels. Instead of tailpipes creating dispersed emissions, the energy (and pollution) can be generated from point sources such as large-scale, centralized facilities with improved efficiency. This allows the possibility of technologies such as carbon sequestration, which are otherwise impossible for mobile applications. Alternatively, distributed energy generation schemes (such as small scale renewable energy sources) can be used, possibly associated with hydrogen stations.

Aside from the energy generation, hydrogen production could be centralized, distributed or a mixture of both. While generating hydrogen at centralized primary energy plants promises higher hydrogen production efficiency, difficulties in high-volume, long range hydrogen transportation (due to factors such as hydrogen damage and the ease of hydrogen diffusion through solid materials) makes electrical energy distribution attractive within a hydrogen economy. In such a scenario, small regional plants or even local filling stations could generate hydrogen using energy provided through the electrical distribution grid. While hydrogen generation efficiency is likely to be lower than for centralized hydrogen generation, losses in hydrogen transport can make such a scheme more efficient in terms of the primary energy used per kilogram of hydrogen delivered to the end user.

The proper balance between hydrogen distribution and long-distance electrical distribution is one of the primary questions that arises in the hydrogen economy.

Again the dilemmas of production sources and transportation of hydrogen can now be overcome using on site (home, business, or fuel station) generation of hydrogen from off grid renewable sources.

Distributed electrolysis

Distributed electrolysis would bypass the problems of distributing hydrogen by distributing electricity instead. It would use existing electrical networks to transport electricity to small, on-site electrolysers located at filling stations. However, accounting for the energy used to produce the electricity and transmission losses will reduce the overall efficiency.

Natural gas combined cycle power plants, which account for almost all builds of new electricity plants in the United States, generate electricity at efficiencies of 60 percent or greater. Increased demand for electricity, whether due to hydrogen cars or other demand, would have the marginal impact of adding new combined cycle power plants. On this basis, distributed production of hydrogen would be roughly 40 percent efficient. However, if the marginal impact is referred to today's power grid, with an efficiency of roughly 40 percent owing to its mix of fuels and conversion methods, the efficiency of distributed hydrogen production would be roughly 25 percent.

The distributed production of hydrogen in this fashion will be expected to generate air emissions of pollutants and carbon dioxide at various points in the supply chain, e.g., electrolysis, transportation and storage. Such externalities as pollution must be weighed against the potential advantages of a hydrogen economy.

Fuel cells as alternative to internal combustion

One of the main offerings of a hydrogen economy is that the fuel can replace the fossil fuel burned in internal combustion engines and turbines as the primary way to convert chemical energy into kinetic or electrical energy; hereby eliminating greenhouse gas emissions and pollution from that engine.

Although hydrogen can be used in conventional internal combustion engines, fuel cells, being electrochemical, have a theoretical efficiency advantage over heat engines. Fuel cells are more expensive to produce than common internal combustion engines, but are becoming cheaper as new technologies and production systems develop.

Some types of fuel cells work with hydrocarbon fuels, while all can be operated on pure hydrogen. In the event that fuel cells become price-competitive with internal combustion engines and turbines, large gas-fired power plants could adopt this technology.

Hydrogen gas must be distinguished as "technical-grade" (five nines pure), which is suitable for applications such as fuel cells, and "commercial-grade", which has carbon- and sulfur-containing impurities, but which can be produced by the much cheaper steam-reformation process. Fuel cells require high purity hydrogen because the impurities would quickly degrade the life of the fuel cell stack.

Much of the interest in the hydrogen economy concept is focused on the use of fuel cells to power electric cars. Current Hydrogen fuel cells suffer from a low power-to-weight ratio, although they store more energy than other electrochemical batteries. Fuel cells are much more efficient than internal combustion engines, and produce no harmful emissions. If a practical method of hydrogen storage is introduced, and fuel cells become cheaper, they can be economically viable to power hybrid fuel cell/battery vehicles, or purely fuel cell-driven ones. The economic viability of fuel cell powered vehicles will improve as the hydrocarbon fuels used in internal combustion engines become more expensive, because of the depletion of easily accessible reserves or economic accounting of environmental impact through such measures as carbon taxes.

Currently it takes 2½ times as much energy to make a hydrogen fuel cell than is obtained from it during its service life.

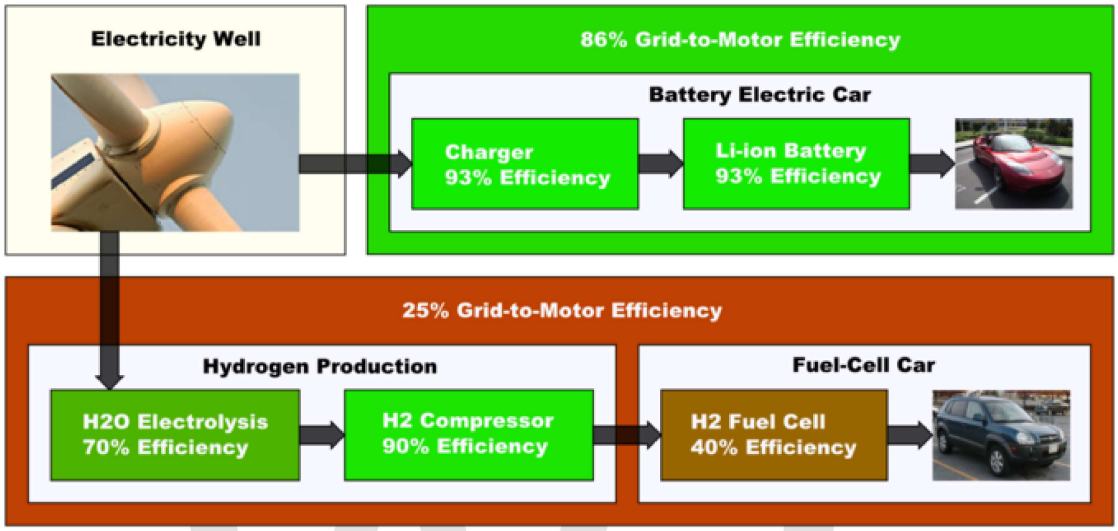
Other fuel cell technologies based on the exchange of metal ions (i.e. zinc-air fuel cells) are typically more efficient at energy conversion than hydrogen fuel cells, but the widespread use of any electrical energy → chemical energy → electrical energy systems would necessitate the production of electricity.

Efficiency as an automotive fuel

Hydrogen has been called one of the least efficient and most expensive possible replacements for gasoline (petrol) in terms of reducing greenhouse gases; other technologies may be less expensive and more quickly implemented. A comprehensive study of hydrogen in transportation applications has found that "there are major hurdles on the path to achieving the vision of the hydrogen economy; the path will not be simple or straightforward". The Ford Motor Company has dropped its plans to develop hydrogen cars, stating that "The next major step in Ford’s plan is to increase over time the volume of electrified vehicles".

An accounting of the energy utilized during a thermodynamic process, known as an energy balance, can be applied to automotive fuels. With today's technology, the manufacture of hydrogen via steam reforming can be accomplished with a thermal efficiency of 75 to 80 percent. Additional energy will be required to liquefy or compress the hydrogen, and to transport it to the filling station via truck or pipeline. The energy that must be utilized per kilogram to produce, transport and deliver hydrogen (i.e., its well-to-tank energy use) is approximately 50 megajoules using technology available in 2004. Subtracting this energy from the enthalpy of one kilogram of hydrogen, which is 141 megajoules, and dividing by the enthalpy, yields a thermal energy efficiency of roughly 60%. Gasoline, by comparison, requires less energy input, per gallon, at the

refinery, and comparatively little energy is required to transport it and store it owing to its high energy density per gallon at ambient temperatures. Well-to-tank, the supply chain for gasoline is roughly 80% efficient (Wang, 2002). The most efficient distribution however is electrical, which is typically 95% efficient. Electric vehicles are typically 3 to 4 times as efficient as hydrogen powered vehicles.



A study of the well-to-wheels efficiency of hydrogen vehicles compared to other vehicles in the Norwegian energy system indicates that hydrogen fuel-cell vehicles tend to be about a third as efficient as EVs when electrolysis is used, with hydrogen Internal Combustion Engines (ICE) being barely a sixth as efficient. Even in the case where hydrogen fuel cells get their hydrogen from natural gas reformation rather than electrolysis, and EVs get their power from a natural gas power plant, the EVs still come out ahead 35% to 25% (and only 13% for a H2 ICE). This compares to 14% for a gasoline ICE, 27% for a gasoline ICE hybrid, and 17% for a diesel ICE, also on a well-to-wheels basis.

Hydrogen safety

Hydrogen has one of the widest explosive/ignition mix range with air of all the gases with few exceptions such as acetylene, silane, and ethylene oxide. That means that whatever the mix proportion between air and hydrogen, a hydrogen leak will most likely lead to an explosion, not a mere flame, when a flame or spark ignites the mixture. This makes the use of hydrogen particularly dangerous in enclosed areas such as tunnels or underground parking. Pure hydrogen-oxygen flames burn in the ultraviolet color range and are nearly invisible to the naked eye, so a flame detector is needed to detect if a hydrogen leak is burning. Hydrogen is odorless and leaks cannot be detected by smell.

Hydrogen codes and standards are codes and standards for hydrogen fuel cell vehicles, stationary fuel cell applications and portable fuel cell applications. There are codes and standards for the safe handling and storage of hydrogen, for example the Standard for the

installation of stationary fuel cell power systems from the National Fire Protection Association.

Codes and standards have repeatedly been identified as a major institutional barrier to deploying hydrogen technologies and developing a hydrogen economy. To enable the commercialization of hydrogen in consumer products, new model building codes and equipment and other technical standards are developed and recognized by federal, state, and local governments.

One of the measures on the roadmap is to implement higher safety standards like early leak detection with hydrogen sensors. The Canadian Hydrogen Safety Program concluded that hydrogen fueling is as safe as, or safer than, CNG fueling. The European Commission has funded the first higher educational program in the world in hydrogen safety engineering at the University of Ulster. It is expected that the general public will be able to use hydrogen technologies in everyday life with at least the same level of safety and comfort as with today's fossil fuels.

Environmental concerns

There are many concerns regarding the environmental effects of the manufacture of hydrogen. Hydrogen is made either by electrolysis of water, or by fossil fuel reforming. Reforming a fossil fuel leads to a higher emissions of carbon dioxide compared with direct use of the fossil fuel in an internal combustion engine. Similarly, if hydrogen is produced by electrolysis from fossil-fuel powered generators, increased carbon dioxide is emitted in comparison with direct use of the fossil fuel.

Using renewable energy source to generate hydrogen by electrolysis would require greater energy input than direct use of the renewable energy to operate electric vehicles, because of the extra conversion stages and losses in distribution.

Like any internal combustion engine, an ICE running on hydrogen may produce nitrous oxides and other pollutants. Air input into the combustion cylinder is approximately 78% nitrogen, and the N_2 molecule has a binding energy of approximately 226 kilocalories per mole. The hydrogen reaction has sufficient energy to break this bond and produce unwanted components such as nitric acid (HNO_3), and hydrogen cyanide gas (HCN), both toxic byproducts. Nitrogen compound emissions from internal combustion engines are a root cause of smog. Hydrogen as transportation fuel, however, is mainly used for fuel cells that do not produce greenhouse gas emission, but water.

There have also been some concerns over possible problems related to hydrogen gas leakage. Molecular hydrogen leaks slowly from most containment vessels. It has been hypothesized that if significant amounts of hydrogen gas (H_2) escape, hydrogen gas may, because of ultraviolet radiation, form free radicals (H) in the stratosphere. These free radicals would then be able to act as catalysts for ozone depletion. A large enough increase in stratospheric hydrogen from leaked H_2 could exacerbate the depletion process. However, the effect of these leakage problems may not be significant. The amount of

hydrogen that leaks today is much lower (by a factor of 10–100) than the estimated 10–20% figure conjectured by some researchers; for example, in Germany, the leakage rate is only 0.1% (less than the natural gas leak rate of 0.7%). At most, such leakage would likely be no more than 1–2% even with widespread hydrogen use, using present technology.

Costs

When evaluating costs, fossil fuels are generally used as the cheapest reference. The energy content of these fuels is not a product of human effort and so has no cost assigned to it. Only the extraction, refining, transportation and production costs are considered. On the other hand, the energy content of a unit of hydrogen fuel must be manufactured, and so has a significant cost, on top of all the costs of refining, transportation, and distribution. Systems which use renewably generated electricity more directly, for example in trolleybuses, or in battery electric vehicles may have a significant economic advantage because there are fewer conversion processes required between primary energy source and point of use.

The barrier to lowering the price of high purity hydrogen is cost of more than 35 kWh of electricity used to generate each kilogram of hydrogen gas.

Demonstrated advances in electrolyzer and fuel cell technology by ITM Power are claimed to have made significant in-roads into addressing the cost of electrolysing water to make hydrogen, making cost effective production of hydrogen from off-grid renewable sources (compared to hydrocarbon fuels) possible for refueling transport and applications for short range business and residential use.

Hydrogen pipelines are more expensive than even long-distance electric lines. Hydrogen is about three times bulkier in volume than natural gas for the same enthalpy, and hydrogen accelerates the cracking of steel (hydrogen embrittlement), which increases maintenance costs, leakage rates, and material costs. The difference in cost is likely to expand with newer technology: wires suspended in air can utilize higher voltage with only marginally increased material costs, but higher pressure pipes require proportionally more material.

Setting up a hydrogen economy would require huge investments in the infrastructure to store and distribute hydrogen to vehicles. In contrast, battery electric vehicles, which are already publicly available, would not necessitate immediate expansion of the existing infrastructure for electricity transmission and distribution, since much of the electricity currently being generated by power plants goes unused at night when the majority of electric vehicles would be recharged. A study conducted by the Pacific Northwest National Laboratory for the US Department of Energy in December 2006 found that the idle off-peak grid capacity in the US would be sufficient to power 84% of all vehicles in the US if they all were immediately replaced with electric vehicles.

Different production methods each have differing associated investment and marginal costs. The energy and feedstock could originate from a multitude of sources i.e. natural gas, nuclear, solar, wind, biomass, coal, other fossil fuels, and geothermal.

Natural Gas at Small Scale

Uses steam reformation. Requires 15.9 million cubic feet (450,000 m^3) of gas, which, if produced by small 500 kg/day reformers at the point of dispensing (i.e., the filling station), would equate to 777,000 reformers costing $1 trillion dollars and producing 150 million tons of hydrogen gas annually. Obviates the need for distribution infrastructure dedicated to hydrogen. $3.00 per GGE (Gallons of Gasoline Equivalent)

Nuclear

Provides energy for electrolysis of water. Would require 240,000 tons of unenriched uranium — that's 2,000 600-megawatt power plants, which would cost $840 billion, or about $2.50 per GGE.

Solar

Provides energy for electrolysis of water. Would require 2,500 kWh of sun per square meter, 113 million 40-kilowatt systems, which would cost $22 trillion, or about $9.50 per GGE.

Wind

Provides energy for electrolysis of water. At 7 meters per second average wind speed, it would require 1 million 2-MW wind turbines, which would cost $3 trillion dollars, or about $3.00 per GGE.

Biomass

Gasification plants would produce gas with steam reformation. 1.5 billion tons of dry biomass, 3,300 plants which would require 113.4 million acres (460,000 km^2) of farm to produce the biomass. $565 billion dollars in cost, or about $1.90 per GGE

Coal

FutureGen plants use coal gasification then steam reformation. Requires 1 billion tons of coal or about 1,000 275-megawatt plants with a cost of about $500 billion, or about $1 per GGE.

Examples and pilot programs

A Mercedes-Benz O530 Citaro powered by hydrogen fuel cells, in Brno, Czech Republic

Several domestic U.S. automobile manufactures have committed to develop vehicles using hydrogen. The distribution of hydrogen for the purpose of transportation is currently being tested around the world, particularly in Portugal, Iceland, Norway, Denmark, Germany, California, Japan and Canada, but the cost is very high.

Some hospitals have installed combined electrolyzer-storage-fuel cell units for local emergency power. These are advantageous for emergency use because of their low maintenance requirement and ease of location compared to internal combustion driven generators.

Iceland has committed to becoming the world's first hydrogen economy by the year 2050. Iceland is in a unique position. Presently, it imports all the petroleum products necessary to power its automobiles and fishing fleet. Iceland has large geothermal resources, so much that the local price of electricity actually is *lower* than the price of the hydrocarbons that could be used to produce that electricity.

Iceland already converts its surplus electricity into exportable goods and hydrocarbon replacements. In 2002, it produced 2,000 tons of hydrogen gas by electrolysis—primarily for the production of ammonia (NH_3) for fertilizer. Ammonia is produced, transported, and used throughout the world, and 90% of the cost of ammonia is the cost of the energy to produce it. Iceland is also developing an aluminium -smelting industry. Aluminium

costs are primarily driven by the cost of the electricity to run the smelters. Either of these industries could effectively export all of Iceland's potential geothermal electricity.

Neither industry directly replaces hydrocarbons. Reykjavík, Iceland, had a small pilot fleet of city buses running on compressed hydrogen, and research on powering the nation's fishing fleet with hydrogen is under way. For more practical purposes, Iceland might process imported oil with hydrogen to extend it, rather than to replace it altogether.

The Reykjavík buses are part of a larger program, HyFLEET: CUTE, operating hydrogen fueled buses in eight European cities. HyFLEET: CUTE buses also operate in Beijing and Perth (see below).

A pilot project demonstrating a hydrogen economy is operational on the Norwegian island of Utsira. The installation combines wind power and hydrogen power. In periods when there is surplus wind energy, the excess power is used for generating hydrogen by electrolysis. The hydrogen is stored, and is available for power generation in periods when there is little wind.

A joint venture between NREL and Xcel Energy is combining wind power and hydrogen power in the same way in Colorado.

Hydro in Newfoundland and Labrador are converting the current wind-diesel Power System on the remote island of Ramea into a Wind-Hydrogen Hybrid Power Systems facility.

A similar pilot project on Stuart Island uses solar power, instead of wind power, to generate electricity. When excess electricity is available after the batteries are full, hydrogen is generated by electrolysis and stored for later production of electricity by fuel cell.

The UK started a fuel cell pilot program in January 2004, the program ran two Fuel cell buses on route 25 in London until December 2005, and switched to route RV1 until January 2007.

The Hydrogen Expedition is currently working to create a hydrogen fuel cell-powered ship and using it to circumnavigate the globe, as a way to demonstrate the capability of hydrogen fuel cells.

Western Australia's Department of Planning and Infrastructure currently operates three Daimler Chrysler Citaro fuel cell buses as part of its Sustainable Transport Energy for Perth Fuel Cells Bus Trial in Perth. The buses are operated by Path Transit on regular Transperth public bus routes. The trial began in September 2004 and concluded in September 2006. The buses' fuel cells use a proton exchange membrane system and are supplied with raw hydrogen from a BP refinery in Kwinana, south of Perth. The hydrogen is a byproduct of the refinery's industrial process. The buses are refueled at a station in the northern Perth suburb of Malaga.

The United Nations Industrial Development Organization (UNIDO) and the Turkish Ministry of Energy and Natural Resources have signed in 2003 a $40M Trust Fund Agreement for the creation in Istanbul of the International Centre for Hydrogen Energy Technologies (UNIDO-ICHET), which started operation in 2004. A hydrogen forklift, a hydrogen cart and a mobile house powered by renewable energies are being demonstrated in UNIDO-ICHET's premises. An uninterruptible power supply system has been has been working since April 2009 in the headquarters of Istanbul Sea Buses company.

Hydrogen-using alternatives to a fully distributive hydrogen economy

Hydrogen is simply a method to store and transmit energy. Various alternative energy transmission and storage scenarios which begin with hydrogen production, but do not use it for all parts of the store and transmission infrastructure, may be more economic, in both near and far term. These include:

Ammonia economy

An alternative to gaseous hydrogen as an energy carrier is to bond it with nitrogen from the air to produce ammonia, which can be easily liquefied, transported, and used (directly or indirectly) as a clean and renewable fuel. The toxicity of ammonia is one of the main issues holding back an ammonia economy.

Hydrogen production of greenhouse-neutral alcohol

The methanol economy is a synfuel production energy plan which may begin with hydrogen production. Hydrogen in a full "hydrogen economy" was initially suggested as a way to make renewable energy in non-polluting form, available to automobiles which are not all-electric. However, a theoretical alternative to direct elemental hydrogen use in vehicles would address the same problem by using centrally produced hydrogen immediately, to make liquid fuels from a CO_2 source. Thus, hydrogen would be used captively to make fuel, and would not require expensive hydrogen transportation or storage. To be greenhouse-neutral, the source for CO_2 in such a plan would need to be from air, biomass, or from CO_2 which would otherwise be scheduled to be released into the air from non-carbon-capture fuel-burning power plants (of which there are likely to be many in the future, since economic carbon capture and storage is site-dependent and difficult to retrofit).

Captive hydrogen production to make more easily transportable and storable transportation fuels (such as alcohols or methane), using CO_2 input, can thus be seen as the artificial, or "non-biological green" analogue of biomass, biodiesel, and vegetable oil technologies. Green plants, in a sense, already use solar power to make captively produced hydrogen, which is then used to make easier-to-store-and-use fuels. In the plant leaf, solar energy is used to split water into hydrogen and oxygen, the latter gas being released. The hydrogen produced is then used "on-site" by the plant to reduce CO_2 from the air into various fuels, such as the cellulose in wood, and the seed oils which are the

basis for vegetable oil, biodiesel, etc. Hydrogen-produced alcohols would thus act as a very similar, but non-biological greenhouse-neutral way of producing energy stores and carriers from locally produced hydrogen (solar or otherwise). By not requiring hydrogen to be produced entirely by plant leaves, they would save cropland. The fuels, however, would be used for purposes of transportation exactly as in plans to use "green fuels." Rather than be transported from its production site, hydrogen in such plans would instead be used centrally and immediately, to produce renewable liquid fuels which may be cycled into the present transportation infrastructure directly, requiring almost no infrastructure change. Moreover, methanol fuel cells are beginning to be demonstrated, so methanol may eventually compete directly with hydrogen in the fuel cell and hybrid market.

The electrical grid plus synthetic methanol fuel cells

Many of the hybrid strategies described above, using captive hydrogen to generate other more easily usable fuels, might be more effective than hydrogen-production alone. Short term energy storage (meaning the energy is used not long after it has been captured) may be best accomplished with battery or even ultracapacitor storage. Longer term energy storage (meaning the energy is used weeks or months after capture) may be better done with synthetic methane or alcohols, which can be stored indefinitely at relatively low cost, and even used directly in some type of fuel cells, for electric vehicles. These strategies dovetail well with the recent interest in Plug-in Hybrid Electric Vehicles, or PHEVs, which use a hybrid strategy of electrical and fuel storage for their energy needs. Hydrogen storage has been proposed by some to be optimal in a narrow range of energy storage time, probably somewhere between a few days and a few weeks. This range is subject to further narrowing with any improvements in battery technology. It is always possible that some kind of breakthrough in hydrogen storage or generation could occur, but this is unlikely given the physical and chemical limitations of the technical choices are fairly well understood.

Captive hydrogen synthetic methane production

In a similar way as with synthetic alcohol production, hydrogen can be used on-site to directly (nonbiologically) produce greenhouse-neutral gaseous fuels. Thus, captive-hydrogen-mediated production of greenhouse-neutral methane has been proposed (note that this is the reverse of the present method of acquiring hydrogen from natural methane, but one that does not require ultimate burning and release of fossil fuel carbon). Captive hydrogen (and carbon dioxide) may be used onsite to *synthesize* methane, using the Sabatier reaction. This process is about 80% efficient, reducing the round trip efficiency to about 20 to 30%, depending on the method of fuel utilization. This is even lower than hydrogen, but the storage costs drop by at least a factor of 3, because of methane's higher boiling point and higher energy density. Liquid methane has 3.2 times the energy density of liquid hydrogen and is easier to store. Additionally, the pipe infrastructure (natural gas pipelines) are already in place. Natural-gas-powered vehicles already exist, and are known to be easier to adapt from existing internal engine technology, than internal combustion autos running directly on hydrogen. Experience with natural gas powered

vehicles shows that methane storage is inexpensive, once one has accepted the cost of conversion to store the fuel. However, the cost of alcohol storage is even lower, so this technology would need to produce methane at a considerable savings with regard to alcohol production. Ultimate mature prices of fuels in the competing technologies are not presently known, but both are expected to offer substantial infrastructural savings over attempts to transport and use hydrogen directly.

Chapter 6

Electric Double-layer Capacitor

Maxwell Technologies "MC" and "BC" series supercapacitors (up to 3000 farad capacitance)

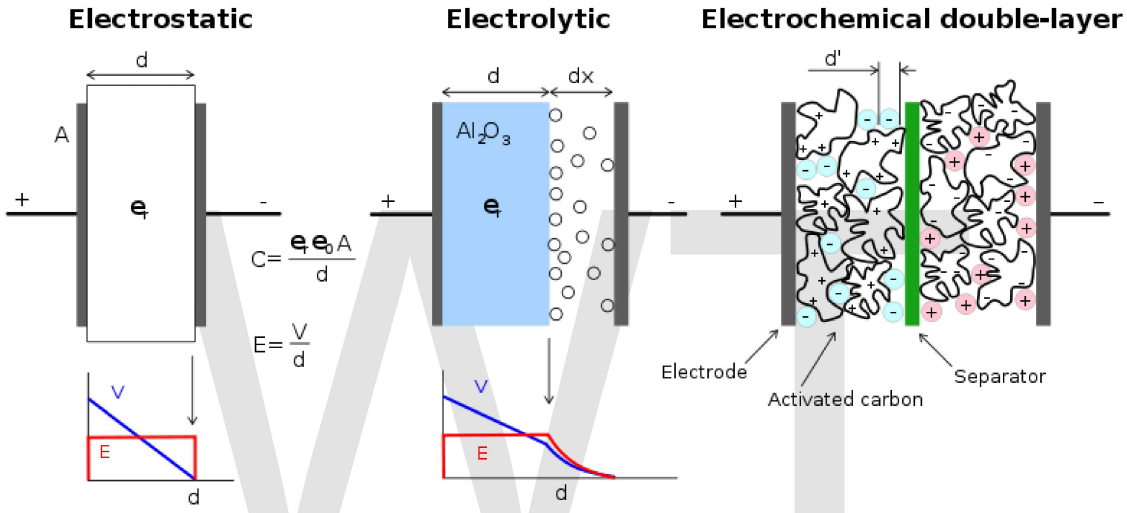
An **electric double-layer capacitor** (**EDLC**), also known as **supercapacitor**, **supercondenser**, **pseudocapacitor**, **electrochemical double layer capacitor**, or **ultracapacitor**, is an electrochemical capacitor with relatively high energy density. Compared to conventional electrolytic capacitors the energy density is typically on the order of thousands of times greater. In comparison with conventional batteries or fuel cells, EDLCs also have a much higher power density.

A typical D-cell sized electrolytic capacitor displays capacitance in the range of tens of millifarads. The same size EDLC might reach several farads, an improvement of two

orders of magnitude. EDLCs usually yield a lower working voltage; as of 2010 larger double-layer capacitors have capacities up to 5,000 farads. Also in 2010, the highest available EDLC energy density is 30 Wh/kg (although 85 Wh/kg has been achieved at room temperature in the lab), lower than rapid-charging lithium-titanate batteries.

EDLCs have a variety of commercial applications, notably in "energy smoothing" and momentary-load devices. They have applications as energy-storage devices used in vehicles, and for smaller applications like home solar energy systems where extremely fast charging is a valuable feature.

Concept



Comparison of construction diagrams of three capacitors. Left: "normal" capacitor, middle: electrolytic, right: electric double-layer capacitor

In a conventional capacitor, energy is stored by the removal of charge carriers, typically electrons, from one metal plate and depositing them on another. This charge separation creates a potential between the two plates, which can be harnessed in an external circuit. The total energy stored in this fashion is proportional to both the amount of charge stored and the potential between the plates. The amount of charge stored per unit voltage is essentially a function of the size, the distance, and the material properties of the plates and the material in between the plates (the dielectric), while the potential between the plates is limited by breakdown of the dielectric. The dielectric controls the capacitor's voltage. Optimizing the material leads to higher energy density for a given size of capacitor.

EDLCs do not have a conventional dielectric. Rather than two separate plates separated by an intervening substance, these capacitors use "plates" that are in fact two layers of the same substrate, and their electrical properties, the so-called "electrical double layer", result in the effective separation of charge despite the vanishingly thin (on the order of nanometers) physical separation of the layers. The lack of need for a bulky layer of

dielectric permits the packing of plates with much larger surface area into a given size, resulting in high capacitances in practical-sized packages.

In an electrical double layer, each layer by itself is quite conductive, but the physics at the interface where the layers are effectively in contact means that no significant current can flow between the layers. However, the double layer can withstand only a low voltage, which means that electric double-layer capacitors rated for higher voltages must be made of matched series-connected individual EDLCs, much like series-connected cells in higher-voltage batteries.

EDLCs have much higher power density than batteries. Power density combines the energy density with the speed that the energy can be delivered to the load. Batteries, which are based on the movement of charge carriers in a liquid electrolyte, have relatively slow charge and discharge times. Capacitors, on the other hand, can be charged or discharged at a rate that is typically limited by current heating of the electrodes. So while existing EDLCs have *energy* densities that are perhaps 1/10th that of a conventional battery, their *power* density is generally 10 to 100 times as great.

History

General Electric engineers experimenting with devices using porous carbon electrodes first observed the EDLC effect in 1957. They believed that the energy was stored in the carbon pores and the device exhibited "exceptionally high capacitance", although the mechanism was unknown at that time.

General Electric did not immediately follow up on this work. In 1966 researchers at Standard Oil of Ohio developed the modern version of the devices, after they accidentally re-discovered the effect while working on experimental fuel cell designs. Their cell design used two layers of activated charcoal separated by a thin porous insulator, and this basic mechanical design remains the basis of most electric double-layer capacitors.

Standard Oil also failed to commercialize their invention, licensing the technology to NEC, who finally marketed the results as “supercapacitors” in 1978, to provide backup power for maintaining computer memory. The market expanded slowly for a time, but starting around the mid-1990s various advances in materials science and refinement of the existing systems led to rapidly improving performance and an equally rapid reduction in cost.

The first trials of supercapacitors in industrial applications were carried out for supporting the energy supply to robots.

In 2005 aerospace systems and controls company Diehl Luftfahrt Elektronik GmbH chose supercapacitors to power emergency actuation systems for doors and evacuation slides in airliners, including the new Airbus 380 jumbo jet. In 2005, the ultracapacitor market was between US $272 million and $400 million, depending on the source.

As of 2007 all solid state micrometer-scale electric double-layer capacitors based on advanced superionic conductors had been for low-voltage electronics such as deep-sub-voltage nanoelectronics and related technologies (the 22 nm technological node of CMOS and beyond).

Comparisons

Supercapacitors have several disadvantages and advantages relative to batteries, as described below.

Disadvantages

- The amount of energy stored per unit weight is generally lower than that of an electrochemical battery (3–5 W·h/kg for an standard ultracapacitor, although 85 W.h/kg has been achieved in the lab as of 2010 compared to 30-40 W·h/kg for a lead acid battery), and about 1/1,000th the volumetric energy density of gasoline.
- Typical of any capacitor, the voltage varies with the energy stored. Effective storage and recovery of energy requires complex electronic control and switching equipment, with consequent energy loss
- Has the highest dielectric absorption of any type of capacitor.
- High self-discharge - the rate is considerably higher than that of an electrochemical battery.
- Cells hold low voltages - serial connections are needed to obtain higher voltages. Voltage balancing is required if more than three capacitors are connected in series.
- Linear discharge voltage prevents use of the full energy spectrum.
- Due to rapid and large release of energy (albeit over short times), EDLC's have the potential to be deadly to humans.

Advantages

- Long life, with little degradation over hundreds of thousands of charge cycles. Due to the capacitor's high number of charge-discharge cycles (millions or more compared to 200 to 1000 for most commercially available rechargeable batteries) it will last for the entire lifetime of most devices, which makes the device environmentally friendly. Rechargeable batteries wear out typically over a few years, and their highly reactive chemical electrolytes present a disposal and safety hazard. Battery lifetime can be optimised by charging only under favorable conditions, at an ideal rate and, for some chemistries, as infrequently as possible. EDLCs can help in conjunction with batteries by acting as a charge conditioner, storing energy from other sources for load balancing purposes and then using any excess energy to charge the batteries at a suitable time.
- Low cost *per cycle*
- Good reversibility
- Very high rates of charge and discharge.

- Extremely low internal resistance (ESR) and consequent high cycle efficiency (95% or more) and extremely low heating levels
- High output power
- High specific power. According to ITS (Institute of Transportation Studies, Davis, California) test results, the specific power of electric double-layer capacitors can exceed 6 kW/kg at 95% efficiency
- Improved safety, no corrosive electrolyte and low toxicity of materials.
- Simple charge methods—no full-charge detection is needed; no danger of overcharging.

Materials

In general, EDLCs improve storage density through the use of a nanoporous material, typically activated charcoal, in place of the conventional insulating barrier. Activated charcoal is a powder made up of extremely small and very "rough" particles, which, in bulk, form a low-density heap with many holes that resembles a sponge. The overall surface area of even a thin layer of such a material is many times greater than a traditional material like aluminum, allowing many more charge carriers (ions or radicals from the electrolyte) to be stored in any given volume. The charcoal, which is not a good insulator, replaces the excellent insulators used in conventional devices, so in general EDLCs can only use low potentials on the order of 2 to 3 V.

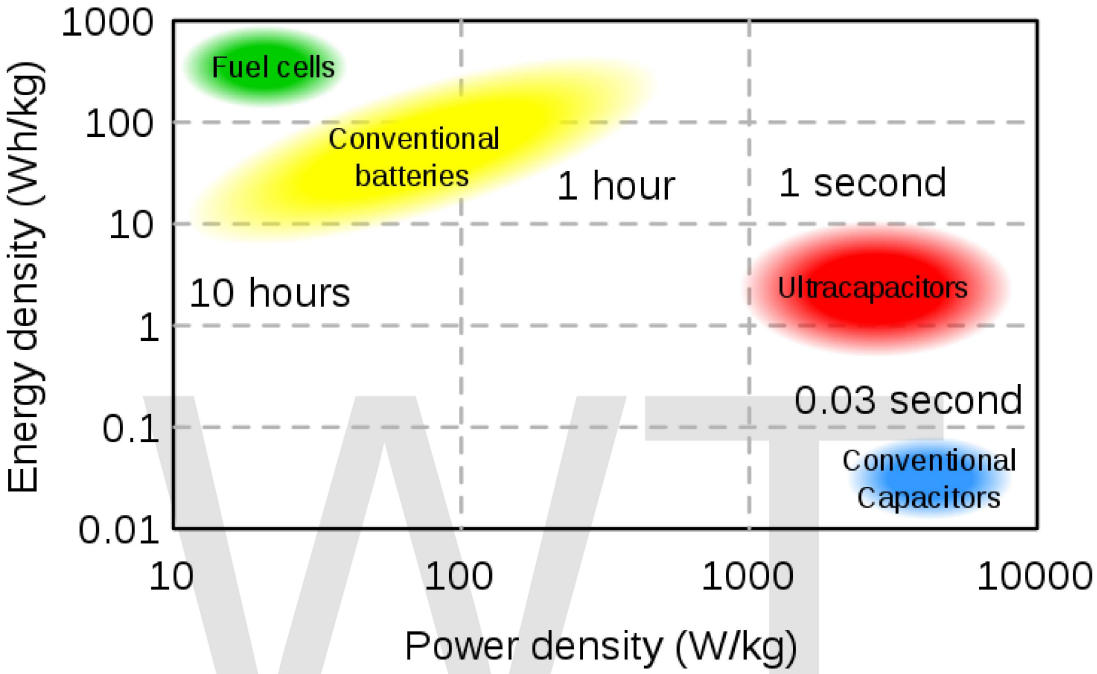
Activated charcoal is not the "perfect" material for this application. The charge carriers are actually (in effect) quite large—especially when surrounded by solventmolecules—and are often larger than the holes left in the charcoal, which are too small to accept them, limiting the storage.

As of 2010 virtually all commercial supercapacitors use powdered activated carbon made from coconut shells. Higher performance devices are available, at a significant cost increase, based on synthetic carbon precursors that are activated with potassium hydroxide (KOH).

Research in EDLCs focuses on improved materials that offer higher *usable* surface areas.

- Graphene has excellent surface area per unit of gravimetric or volumetric densities, is highly conductive and can now be produced in various labs, but is not available in production quantities. Specific energy density of 85.6 Wh/kg at room temperature and 136 Wh/kg at 80 °C (all based on the total electrode weight), measured at a current density of 1 A/g have been observed. These energy density values are comparable to that of the Nickel metal hydride battery. The device makes full utilization of the highest intrinsic surface capacitance and specific surface area of single-layer graphene by preparing curved graphene sheets that do not restack face-to-face. The curved shape enables the formation of mesopores accessible to and wettable by environmentally benign ionic liquids capable of operating at a voltage >4 V.

- Carbon nanotubes have excellent nanoporosity properties, allowing tiny spaces for the polymer to sit in the tube and act as a dielectric. Carbon nanotubes can store about the same charge as charcoal (which is almost pure carbon) per unit surface area but nanotubes can be arranged in a more regular pattern that exposes greater suitable surface area.



Ragone chart showing energy density vs.power density for various energy-storage devices

- Some polymers (eg. polyacenes) have a redox (reduction-oxidation) storage mechanism along with a high surface area.

- Carbon aerogel provides extremely high surface area gravimetric densities of about 400–1000 m^2/g. The electrodes of aerogel supercapacitors are a composite material usually made of non-woven paper made from carbon fibers and coated with organic aerogel, which then undergoes pyrolysis. The carbon fibers provide structural integrity and the aerogel provides the required large surface area. Small aerogel supercapacitors are being used as backup electricity storage in microelectronics. Aerogel capacitors can only work at a few volts; higher voltages ionize the carbon and damage the capacitor. Carbon aerogel capacitors have achieved 325 J/g (90 W·h/kg) energy density and 20 W/g power density.

- Solid activated carbon, also termed *consolidated amorphous carbon* (CAC). It can have a surface area exceeding 2800 m^2/g and may be cheaper to produce than aerogel carbon.

- Tunable nanoporous carbon exhibits systematic pore size control. H_2 adsorption treatment can be used to increase the energy density by as much as 75% over what was commercially available as of 2005.

- Mineral-based carbon is a nonactivated carbon, synthesised from metal or metalloid carbides, e.g. SiC, TiC, Al_4C_3. The synthesised nanostructured porous carbon, often called Carbide Derived Carbon (CDC), has a surface area of about 400 m^2/g to 2000 m^2/g with a specific capacitance of up to 100 F/mL (in organic electrolyte). As of 2006 this material was used in a supercapacitor with a volume of 135 mL and 200 g weight having 1.6 kF capacitance. The energy density is more than 47 kJ/L at 2.85 V and power density of over 20 W/g.

- In August 2007 researchers combined a biodegradable paper battery with aligned carbon nanotubes, designed to function as both a lithium-ion battery and a supercapacitor (called *bacitor*). The device employed an ionic liquid, essentially a liquid salt, as the electrolyte. The paper sheets can be rolled, twisted, folded, or cut with no loss of integrity or efficiency, or stacked, like ordinary paper (or a voltaic pile), to boost total output. They can be made in a variety of sizes, from postage stamp to broadsheet. Their light weight and low cost make them attractive for portable electronics, aircraft, automobiles, and toys (such as model aircraft), while their ability to use electrolytes in blood make them potentially useful for medical devices such as pacemakers.

- Other teams are experimenting with custom materials made of activated polypyrrole, and nanotube-impregnated papers.

Density

The energy density of existing commercial EDLCs ranges from around 0.5 to 30 W·h/kg including lithium ion capacitors, known also as a "hybrid capacitor". Experimental electric double-layer capacitors have demonstrated densities of 30 W·h/kg and have been shown to be scalable to at least 136 W·h/kg, while others expect to offer energy densities of about 400 W·h/kg. For comparison, a conventional lead-acid battery stores typically 30 to 40 W·h/kg and modern lithium-ion batteries about 160 W·h/kg. Gasoline has a net calorific value (NCV) of around 12,000 W·h/kg; automobile applications operate at about 20% tank-to-wheel efficiency, giving an effective energy density of 2,400 W·h/kg.

Applications

Vehicles

Heavy and public transport

Some of the earliest uses were motor startup capacitors for large engines in tanks and submarines, and as the cost has fallen they have started to appear on diesel trucks and railroad locomotives. In the 00's they attracted attention in the green energy world, where

their ability to charge much faster than batteries makes them particularly suitable for regenerative braking applications. New technology in development could potentially make EDLCs with high enough energy density to be an attractive replacement for batteries in all-electric cars and plug-in hybrids, as EDLCs charge quickly and are stable with respect to temperature.

China is experimenting with a new form of electric bus (capabus) that runs without powerlines using large onboard EDLCs, which quickly recharge whenever the bus is at any bus stop (under so-called **electric umbrellas**), and fully charge in the terminus. A few prototypes were being tested in Shanghai in early 2005. In 2006, two commercial bus routes began to use electric double-layer capacitor buses; one of them is route 11 in Shanghai.

In 2001 and 2002 VAG, the public transport operator in Nuremberg, Germany tested an hybrid bus that uses a diesel-electric battery drive system with electric double-layer capacitors. Since 2003 Mannheim Stadtbahn in Mannheim, Germany has operated a light-rail vehicle (LRV) that uses EDLCs to store braking energy.

Other public transport manufacturers are developing EDLC technology, including mobile storage and a stationary trackside power supply.

A triple hybrid forklift truck uses fuel cells and batteries as primary energy storage and EDLCs to supplement this energy storage solution.

Automotive

Ultracapacitors are used in some concept prototype vehicles, in order to keep batteries within resistive heating limits and extend battery life. The ultrabattery combines a supercapacitor and a battery in one unit, creating an electric vehicle battery that lasts longer, costs less and is more powerful than current plug-in hybrid electric vehicles (PHEVs).

Motor racing

The FIA, the governing body for many motor racing events, proposed in the *Power-Train Regulation Framework for Formula 1* version 1.3 of 23 May 2007 that a new set of power train regulations be issued that includes a hybrid drive of up to 200 kW input and output power using "superbatteries" made with both batteries and supercapacitors.

Consumer electronics

EDLCs can be used in PC Cards, flash photography devices in digital cameras, flashlights, portable media players, and in automated meter reading, particularly where extremely fast charging is desirable.

In 2007, a cordless electric screwdriver that uses an EDLC for energy storage was produced. It charges in 90 seconds, retains 85% of the charge after 3 months, and holds enough charge for about half the screws (22) a comparable screwdriver with a rechargeable battery will handle (37). Two LED flashlights using EDLCs were released in 2009. They charge in 90 seconds.

Alternative energy

The idea of replacing batteries with capacitors in conjunction with novel energy sources became a conceptual umbrella of the Green Electricity (GEL) Initiative, introduced by Dr. Alexander Bell. One successful GEL Initiative concept was a muscle-driven autonomous solution that employs a multi-farad EDLC as energy storage to power a variety of portable electrical and electronic devices such as MP3 players, AM/FM radios, flashlights, cell phones, and emergency kits.

Price

Costs have fallen quickly, with cost per kilojoule dropping faster than cost per farad. As of 2006 the cost of supercapacitors was 1 cent per farad and $2.85 per kilojoule, and was expected to drop further.

Market

According to Innovative Research and Products (iRAP), ultracapacitor market growth will continue during 2009 to 2014. Worldwide business, over US$275 million in 2009, will continue to grow at an AAGR of 21.4% through 2014.

Chapter 7

Machine Translation

Machine translation, sometimes referred to by the abbreviation **MT**, also called **computer-aided translation**, **machine-aided human translation MAHT** and **interactive translation**, is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another.

At its basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus and statistical techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.

Current machine translation software often allows for customisation by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than conversation or less standardised text.

Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators and, in a very limited number of cases, can even produce output that can be used as is (e.g., weather reports).

The progress and potential of machine translation has been debated much through its history. Since the 1950s, a number of scholars has questioned the possibility of achieving fully automatic machine translation of high quality. Some critics claim that there are in-principle obstacles to automatizing the translation process.

History

The idea of machine translation may be traced back to the 17th century. In 1629, René Descartes proposed a universal language, with equivalent ideas in different tongues sharing one symbol. In the 1950s, The Georgetown experiment (1954) involved fully-

automatic translation of over sixty Russian sentences into English. The experiment was a great success and ushered in an era of substantial funding for machine-translation research. The authors claimed that within three to five years, machine translation would be a solved problem.

Real progress was much slower, however, and after the ALPAC report (1966), which found that the ten-year-long research had failed to fulfill expectations, funding was greatly reduced. Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for machine translation.

The idea of using digital computers for translation of natural languages was proposed as early as 1946 by A. D. Booth and possibly others. The Georgetown experiment was by no means the first such application, and a demonstration was made in 1954 on the APEXC machine at Birkbeck College (University of London) of a rudimentary translation of English into French. Several papers on the topic were published at the time, and even articles in popular journals. A similar application, also pioneered at Birkbeck College at the time, was reading and composing Braille texts by computer.

Translation process

The translation process may be stated as:

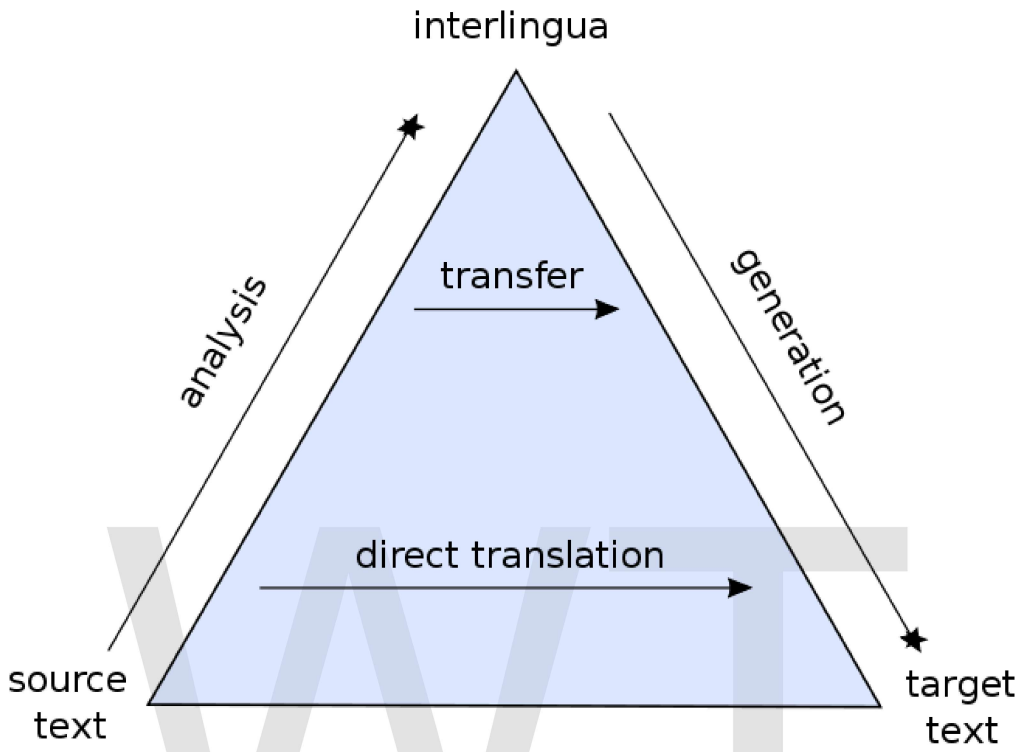
1. Decoding the meaning of the source text; and
2. Re-encoding this meaning in the target language.

Behind this ostensibly simple procedure lies a complex cognitive operation. To decode the meaning of the source text in its entirety, the translator must interpret and analyse all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language, as well as the culture of its speakers. The translator needs the same in-depth knowledge to re-encode the meaning in the target language.

Therein lies the challenge in machine translation: how to program a computer that will "understand" a text as a person does, and that will "create" a new text in the target language that "sounds" as if it has been written by a person.

This problem may be approached in a number of ways.

Approaches



Pyramid showing comparative depths of intermediary representation, interlingual machine translation at the peak, followed by transfer-based, then direct translation.

Machine translation can use a method based on linguistic rules, which means that words will be translated in a linguistic way — the most suitable (orally speaking) words of the target language will replace the ones in the source language.

It is often argued that the success of machine translation requires the problem of natural language understanding to be solved first.

Generally, rule-based methods parse a text, usually creating an intermediary, symbolic representation, from which the text in the target language is generated. According to the nature of the intermediary representation, an approach is described as interlingual machine translation or transfer-based machine translation. These methods require extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules.

Given enough data, machine translation programs often work well enough for a native speaker of one language to get the approximate meaning of what is written by the other native speaker. The difficulty is getting enough data of the right kind to support the

particular method. For example, the large multilingual corpus of data needed for statistical methods to work is not necessary for the grammar-based methods. But then, the grammar methods need a skilled linguist to carefully design the grammar that they use.

To translate between closely related languages, a technique referred to as shallow-transfer machine translation may be used.

Rule-based

The rule-based machine translation paradigm includes transfer-based machine translation, interlingual machine translation and dictionary-based machine translation paradigms.

Interlingual

Interlingual machine translation is one instance of rule-based machine-translation approaches. In this approach, the source language, i.e. the text to be translated, is transformed into an interlingual, i.e. source-/target-language-independent representation. The target language is then generated out of the interlingua.

Dictionary-based

Machine translation can use a method based on dictionary entries, which means that the words will be translated as they are by a dictionary.

Statistical

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament and EUROPARL, the record of the European Parliament. Where such corpora are available, impressive results can be achieved translating texts of a similar kind, but such corpora are still very rare. The first statistical machine translation software was CANDIDE from IBM. Google used SYSTRAN for several years, but switched to a statistical translation method in October 2007. Recently, they improved their translation capabilities by inputting approximately 200 billion words from United Nations materials to train their system. Accuracy of the translation has improved.

Example-based

Example-based machine translation (EBMT) approach was proposed by Makoto Nagao in 1984. It is often characterised by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning.

Hybrid MT

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies. Several MT companies (Asia Online, LinguaSys, Systran, UPV) are claiming to have a hybrid approach using both rules and statistics. The approaches differ in a number of ways:

- **Rules post-processed by statistics**: Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine.
- **Statistics guided by rules**: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating.

Major issues

Disambiguation

Word-sense disambiguation concerns finding a suitable translation when a word can have more than one meaning. The problem was first raised in the 1950s by Yehoshua Bar-Hillel. He pointed out that without a "universal encyclopedia", a machine would never be able to distinguish between the two meanings of a word. Today there are numerous approaches designed to overcome this problem. They can be approximately divided into "shallow" approaches and "deep" approaches.

Shallow approaches assume no knowledge of the text. They simply apply statistical methods to the words surrounding the ambiguous word. Deep approaches presume a comprehensive knowledge of the word. So far, shallow approaches have been more successful.

The late Claude Piron, a long-time translator for the United Nations and the World Health Organization, wrote that machine translation, at its best, automates the easier part of a translator's job; the harder and more time-consuming part usually involves doing extensive research to resolve ambiguities in the source text, which the grammatical and lexical exigencies of the target language require to be resolved:

> Why does a translator need a whole workday to translate five pages, and not an hour or two? ..... About 90% of an average text corresponds to these simple conditions. But unfortunately, there's the other 10%. It's that part that requires six [more] hours of work. There are ambiguities one has to resolve. For instance, the author of the source text, an Australian physician, cited the example of an epidemic which was declared during World War II in a "Japanese prisoner of war camp". Was he talking about an American camp with Japanese prisoners or a Japanese camp with American prisoners? The English has two senses. It's

necessary therefore to do research, maybe to the extent of a phone call to Australia.

The ideal deep approach would require the translation software to do all the research necessary for this kind of disambiguation on its own; but this would require a higher degree of AI than has yet been attained. A shallow approach which simply guessed at the sense of the ambiguous English phrase that Piron mentions (based, perhaps, on which kind of prisoner-of-war camp is more often mentioned in a given corpus) would have a reasonable chance of guessing wrong fairly often. A shallow approach that involves "ask the user about each ambiguity" would, by Piron's estimate, only automate about 25% of a professional translator's job, leaving the harder 75% still to be done by a human.

Named entities

Related to named entity recognition in information extraction.

Applications

There are now many software programs for translating natural language, several of them online, such as:

- Ta with you is specialized in customized machine translation solutions in any language. Their web-based user interface makes it easy for any Language Service Provider to generate any combination of domain and language pair to achieve the best quality. Their solution works with almost human quality for combinations from/to Spanish.
- LinguaSys provides highly customized hybrid machine translation that can go from any language to any language.
- Asia Online provides a custom machine translation engine building capability that they claim gives near-human quality compared to the "gist" based quality of free online engines. Asia Online also provides tools to edit and create custom machine translation engines with their Language Studio suite of products.
- Hindi to Punjabi Machine Translation System, provides machine translation using a direct approach. It translates Hindi into Punjabi. It also features writing e-mail in the Hindi language and sending the same in Punjabi to the recipient.
- Arabic machine translation in multilingual framework.
- Worldlingo provides machine translation using both statistical based TE's and rule based TE's. Most recognizable as the MT partner in Microsoft Windows and Microsoft Mac Office.
- Power Translator
- SYSTRAN, which powers Yahoo! Babel Fish
- Promt, which powers online translation services at Voila.fr and Orange.fr
- AppTek, which released a hybrid MT system in 2009.
- Toggletext uses a transfer-based system (known as Kataku) to translate between English and Indonesian.

- Anusaaraka A free open source machine translation from English to Hindi based on Panini grammar and uses state of the art NLP tools. Can be used online and downloaded from
- Apertium, a free and open source machine translation platform (WinXLator gives this a Windows GUI, but it is likely to be in violation of the Apertium GPL license)
- Google Translate A free online translator from Google.

Other translation software, most of them running under Microsoft Windows, includes

- Translation memory tools, such as Globalsight, SDL Trados, Wordfast, Deja Vu, Swordfish, and
- localization tools, such and Alchemy CATALYST and Multilizer.
- SiShiTra — A hybrid machine translation engine for Spanish-Catalan translation.

(A comparison test of software of this kind may be seen here.) A number of translation software programs are available free of charge, e.g. ForeignDesk and the multiplatform Okapi Framework and OmegaT+.

While no system provides the holy grail of fully-automatic high-quality machine translation of unrestricted text, many fully-automated systems produce reasonable output. The quality of machine translation is substantially improved if the domain is restricted and controlled.

Despite their inherent limitations, MT programs are used around the world. Probably the largest institutional user is the European Commission. The MOLTO project, for example, coordinated by the University of Gothenburg, received more than 2.375 million euros project support from the EU to create a reliable translation tool that covers a majority of the EU languages.

Google has claimed that promising results were obtained using a proprietary statistical machine translation engine. The statistical translation engine used in the Google language tools for Arabic <-> English and Chinese <-> English had an overall score of 0.4281 over the runner-up IBM's BLEU-4 score of 0.3954 (Summer 2006) in tests conducted by the National Institute for Standards and Technology.

With the recent focus on terrorism, the military sources in the United States have been investing significant amounts of money in natural language engineering. *In-Q-Tel* (a venture capital fund, largely funded by the US Intelligence Community, to stimulate new technologies through private sector entrepreneurs) brought up companies like Language Weaver. Currently the military community is interested in translation and processing of languages like Arabic, Pashto, and Dari. The Information Processing Technology Office in DARPA hosts programs like TIDES and Babylon Translator. US Air Force has awarded a $1 million contract to develop a language translation technology.

The notable rise of social networking on the web in recent years has created yet another niche for the application of machine translation software – in utilities such as Facebook, or instant messaging clients such as Skype, GoogleTalk, MSN Messenger, etc. – allowing users speaking different languages to communicate with each other. Machine translation applications have also been released for most mobile devices, including mobile telephones, pocket PCs, PDAs, etc. Due to their portability, such instruments have come to be designated as mobile translation tools enabling mobile business networking between partners speaking different languages, or facilitating both foreign language learning and unaccompanied traveling to foreign countries without the need of the intermediation of a human translator.

Evaluation

Machine translation systems and output can be evaluated along numerous dimensions. The intended use of the translation, characteristics of the MT software, the nature of the translation process, etc., all affect how one evaluates MT systems and their output.

There are various means for evaluating the output quality of machine translation systems. The oldest is the use of human judges to assess a translation's quality. Even though human evaluation is time-consuming, it is still the most reliable way to compare different systems such as rule-based and statistical systems. Automated means of evaluation include BLEU, NIST and METEOR.

Relying exclusively on unedited machine translation ignores the fact that communication in human language is context-embedded and that it takes a person to comprehend the context of the original text with a reasonable degree of probability. It is certainly true that even purely human-generated translations are prone to error. Therefore, to ensure that a machine-generated translation will be useful to a human being and that publishable-quality translation is achieved, such translations must be reviewed and edited by a human. The late Claude Piron wrote that machine translation, at its best, automates the easier part of a translator's job; the harder and more time-consuming part usually involves doing extensive research to resolve ambiguities in the source text, which the grammatical and lexical exigencies of the target language require to be resolved. Such research is a necessary prelude to the pre-editing necessary in order to provide input for machine-translation software such that the output will not be meaningless.

In certain applications, however, e.g., product descriptions written in a controlled language, a dictionary-based machine-translation system has produced satisfactory translations that require no human intervention save for quality inspection.

Chapter 8

GPGPU

General-purpose computing on graphics processing units (**GPGPU**, also referred to as **GPGP** and less often **GP²**) is the technique of using a GPU, which typically handles computation only for computer graphics, to perform computation in applications traditionally handled by the CPU. It is made possible by the addition of programmable stages and higher precision arithmetic to the rendering pipelines, which allows software developers to use stream processing on non-graphics data.

GPU improvements

GPU functionality has, traditionally, been very limited. In fact, for many years the GPU was only used to accelerate certain parts of the graphics pipeline. Some improvements were needed before GPGPU became feasible.

Programmability

Programmable vertex and fragment shaders were added to the graphics pipeline to enable game programmers to generate even more realistic effects. Vertex shaders allow the programmer to alter per-vertex attributes, such as position, color, texture coordinates, and normal vector. Fragment shaders are used to calculate the color of a fragment, or per-pixel. Programmable fragment shaders allow the programmer to substitute, for example, a lighting model other than those provided by default by the graphics card, typically simple Gouraud shading. Shaders have enabled graphics programmers to create lens effects, displacement mapping, and depth of field.

The programmability of the pipelines have trended, according to Microsoft's DirectX specification, with DirectX8 introducing Shader Model 1.1, DirectX8.1 Pixel Shader Models 1.2, 1.3 and 1.4, and DirectX9 defining Shader Model 2.x and 3.0. Each shader model increased the programming model flexibilities and capabilities, ensuring the conforming hardware follows suit. The DirectX10 specification introduces Shader Model 4.0 which unifies the programming specification for vertex, geometry ("Geometry Shaders" are new to DirectX10) and fragment processing allowing for a better fit for unified shader hardware, thus providing a single computational pool of programmable resource.

Data types

Pre-DirectX9 graphics cards only supported paletted or integral color types. Various formats are available, each containing a red element, a green element, and a blue element. Sometimes an additional alpha value is added, to be used for transparency. Common formats are:

- 8 bits per pixel – Palette mode, where each value is an index in a table with the real color value specified in one of the other formats. Possibly two bits for red, three bits for green, and three bits for blue.
- 16 bits per pixel – Usually allocated as five bits for red, six bits for green, and five bits for blue.
- 24 bits per pixel – eight bits for each of red, green, and blue
- 32 bits per pixel – eight bits for each of red, green, blue, and alpha

For early fixed-function or limited programmability graphics (i.e. up to and including DirectX8.1-compliant GPUs) this was sufficient because this is also the representation used in displays. This representation does have certain limitations, however. Given sufficient graphics processing power even graphics programmers would like to use better formats, such as floating point data formats, in order to obtain effects such as high dynamic range imaging. Many GPGPU applications require floating point accuracy, which came with graphics cards conforming to the DirectX9 specification.

DirectX9 Shader Model 2.x suggested the support of two precision types: full and partial precision. Full precision support could either be FP32 and FP24 (floating point 24-bit per component) or greater, while partial precision was FP16. ATI's R300 series of GPUs supported FP24 precision only in the programmable fragment pipeline (although FP32 was supported in the vertex processors) while Nvidia's NV30 series supported both FP16 and FP32; other vendors such as S3 Graphics and XGI supported a mixture of formats up to FP24.

Shader Model 3.0 altered the specification, increasing full precision requirements to a minimum of FP32 support in the fragment pipeline. ATI's Shader Model 3.0 compliant R5xx generation (Radeon X1000 series) supports just FP32 throughout the pipeline while Nvidia's NV4x and G7x series continued to support both FP32 full precision and FP16 partial precisions. Although not stipulated by Shader Model 3.0, both ATI and Nvidia's Shader Model 3.0 GPUs introduced support for blendable FP16 render targets, more easily facilitating the support for High Dynamic Range Rendering.

The implementations of floating point on Nvidia GPUs are mostly IEEE compliant; however, this is not true across all vendors. This has implications for correctness which are considered important to some scientific applications. While 64-bit floating point values (double precision float) are commonly available on CPUs, these are not universally supported on GPUs; some GPU architectures sacrifice IEEE compliance while others lack double-precision altogether. There have been efforts to emulate double-

precision floating point values on GPUs; however, the speed tradeoff negates any benefit to offloading the computation onto the GPU in the first place.

Most operations on the GPU operate in a vectorized fashion: a single operation can be performed on up to four values at once. For instance, if one color <R1, G1, B1> is to be modulated by another color <R2, G2, B2>, the GPU can produce the resulting color <R1*R2, G1*G2, B1*B2> in a single operation. This functionality is useful in graphics because almost every basic data type is a vector (either 2-, 3-, or 4-dimensional). Examples include vertices, colors, normal vectors, and texture coordinates. Many other applications can put this to good use, and because of their higher performance, vector instructions (SIMD) have long been available on CPUs.

In November 2006 Nvidia launched CUDA, a SDK and API that allows a programmer to use the C programming language to code algorithms for execution on Geforce 8 series GPUs. AMD offers a similar SDK+API for their ATI-based GPUs, that SDK and technology is called FireStream SDK (formerly a thin hardware interface[1]Close to Metal), designed to compete directly with Nvidia's CUDA. OpenCL from Khronos Group is used paired with OpenGL to unify the C languages extension between different architectures; it support both Nvidia and AMD/ATI GPUs and general-purpose CPUs too. GPGPU compared, for example, to traditional floating point accelerators such as the 64-bit CSX700 boards from ClearSpeed that are used in today's supercomputers, current top-end GPUs from Nvidia and AMD emphasize single-precision (32-bit) computation; double-precision (64-bit) computation executes much more slowly.

GPGPU programming concepts

GPUs are designed specifically for graphics and thus are very restrictive in terms of operations and programming. Because of their nature, GPUs are only effective at tackling problems that can be solved using stream processing and the hardware can only be used in certain ways.

Stream processing

GPUs can only process independent vertices and fragments, but can process many of them in parallel. This is especially effective when the programmer wants to process many vertices or fragments in the same way. In this sense, GPUs are stream processors – processors that can operate in parallel by running a single kernel on many records in a stream at once.

A **stream** is simply a set of records that require similar computation. Streams provide data parallelism. **Kernels** are the functions that are applied to each element in the stream. In the GPUs, **vertices** and **fragments** are the elements in streams and vertex and fragment shaders are the kernels to be run on them. Since GPUs process elements independently there is no way to have shared or static data. For each element we can only read from the input, perform operations on it, and write to the output. It is permissible to

have multiple inputs and multiple outputs, but never a piece of memory that is both readable and writable.

Arithmetic intensity is defined as the number of operations performed per word of memory transferred. It is important for GPGPU applications to have high arithmetic intensity else the memory access latency will limit computational speedup.

Ideal GPGPU applications have large data sets, high parallelism, and minimal dependency between data elements.

GPU programming concepts

Computational resources

There are a variety of computational resources available on the GPU:

- Programmable processors – Vertex, primitive, and fragment pipelines allow programmer to perform kernel on streams of data
- Rasterizer – creates fragments and interpolates per-vertex constants such as texture coordinates and color
- Texture Unit – read only memory interface
- Framebuffer – write only memory interface

In fact, the programmer can substitute a write only texture for output instead of the framebuffer. This is accomplished either through Render to Texture (RTT), Render-To-Backbuffer-Copy-To-Texture(RTBCTT), or the more recent stream-out.

Textures as stream

The most common form for a stream to take in GPGPU is a 2D grid because this fits naturally with the rendering model built into GPUs. Many computations naturally map into grids: matrix algebra, image processing, physically based simulation, and so on.

Since textures are used as memory, texture lookups are then used as memory reads. Certain operations can be done automatically by the GPU because of this.

Kernels

Kernels can be thought of as the body of loops. For example, if the programmer were operating on a grid on the CPU they might have code that looked like this:

```
// Input and output grids have 10000 x 10000 or 100 million elements.

void transform_10k_by_10k_grid(float in[10000][10000], float
out[10000][10000])
{
  for(int x = 0; x < 10000; x++)
```

```
  {
    for(int y = 0; y < 10000; y++)
    {
      // The next line is executed 100 million times
      out[x][y] = do_some_hard_work(in[x][y]);
    }
  }
}
```

On the GPU, the programmer only specifies the body of the loop as the kernel and what data to loop over by invoking geometry processing.

Flow control

In sequential code it is possible to control the flow of the program using if-then-else statements and various forms of loops. Such flow control structures have only recently been added to GPUs. Conditional writes could be accomplished using a properly crafted series of arithmetic/bit operations, but looping and conditional branching were not possible.

Recent GPUs allow branching, but usually with a performance penalty. Branching should generally be avoided in inner loops, whether in CPU or GPU code, and various techniques, such as static branch resolution, pre-computation, and Z-cull can be used to achieve branching when hardware support does not exist.

GPU techniques

Map

The map operation simply applies the given function (the kernel) to every element in the stream. A simple example is multiplying each value in the stream by a constant (increasing the brightness of an image). The map operation is simple to implement on the GPU. The programmer generates a fragment for each pixel on screen and applies a fragment program to each one. The result stream of the same size is stored in the output buffer.

Reduce

Some computations require calculating a smaller stream (possibly a stream of only 1 element) from a larger stream. This is called a reduction of the stream. Generally a reduction can be accomplished in multiple steps. The results from the previous step are used as the input for the current step and the range over which the operation is applied is reduced until only one stream element remains.

Stream filtering

Stream filtering is essentially a non-uniform reduction. Filtering involves removing items from the stream based on some criteria.

Scatter

The scatter operation is most naturally defined on the vertex processor. The vertex processor is able to adjust the position of the vertex, which allows the programmer to control where information is deposited on the grid. Other extensions are also possible, such as controlling how large an area the vertex affects.

The fragment processor cannot perform a direct scatter operation because the location of each fragment on the grid is fixed at the time of the fragment's creation and cannot be altered by the programmer. However, a logical scatter operation may sometimes be recast or implemented with an additional gather step. A scatter implementation would first emit both an output value and an output address. An immediately following gather operation uses address comparisons to see whether the output value maps to the current output slot.

Gather

The fragment processor is able to read textures in a random access fashion, so it can gather information from any grid cell, or multiple grid cells, as desired.

Sort

The sort operation transforms an unordered set of elements into an ordered set of elements. The most common implementation on GPUs is using sorting networks.

Search

The search operation allows the programmer to find a particular element within the stream, or possibly find neighbors of a specified element. The GPU is not used to speed up the search for an individual element, but instead is used to run multiple searches in parallel.

Data structures

A variety of data structures can be represented on the GPU:

- Dense arrays
- Sparse arrays – static or dynamic
- Adaptive structures

Applications

The following are some of the areas where GPUs have been used for general purpose computing:

- MATLAB acceleration using the Parallel Computing Toolbox and Distributed Computing Server, as well as 3rd party packages like Jacket.
- k-nearest neighbor algorithm
- Computer clusters or a variation of a parallel computing (utilizing GPU cluster technology) for highly calculation-intensive tasks:
    - High-performance computing clusters (HPC clusters) (often referred to as supercomputers)
        - including cluster technologies like Message Passing Interface, and single-system image (SSI), distributed computing, and Beowulf
    - Grid computing (a form of distributed computing) (networking many heterogeneous computers to create a virtual computer architecture)
    - Load-balancing clusters (sometimes referred to as a server farm)
- Physical based simulation and physics engines (usually based on Newtonian physics models)
    - Conway's Game of Life, cloth simulation, incompressible fluid flow by solution of Navier-Stokes equations
- Statistical physics
    - Ising model
- Lattice gauge theory
- Segmentation – 2D and 3D
- Level-set methods
- CT reconstruction
- Fast Fourier transform
- Tone mapping
- Audio signal processing
    - Audio and Sound Effects Processing, to use a GPU for DSP (digital signal processing)
    - Analog signal processing
    - Speech processing
- Digital image processing
- Video Processing
    - Hardware accelerated video decoding and post-processing
        - Motion compensation (mo comp)
        - Inverse discrete cosine transform (iDCT)
        - Variable-length decoding (VLD)
        - Inverse quantization (IQ)
        - In-loop deblocking
        - Bitstream processing (CAVLC/CABAC) using special purpose hardware for this task because this is a serial task not suitable for regular GPGPU computation
        - Deinterlacing

        - Spatial-temporal de-interlacing
      - Noise reduction
      - Edge enhancement
      - Color correction
    - Hardware accelerated video encoding and pre-processing
- Raytracing
- Global illumination – photon mapping, radiosity, subsurface scattering
- Geometric computing – constructive solid geometry, distance fields, collision detection, transparency computation, shadow generation
- Scientific computing
    - Monte Carlo simulation of light propagation
    - Weather forecasting
    - Climate research
    - Molecular modeling on GPU
    - Quantum mechanical physics
    - Astrophysics
- Bioinformatics
- Computational finance
- Medical imaging
- Computer vision
- Digital signal processing / signal processing
- Control engineering
- Neural networks
- Database operations
- Lattice Boltzmann methods
- Cryptography and cryptanalysis
    - Implementation of MD6
    - Implementation of AES
    - Implementation of DES
    - Implementation of RSA
    - Implementation of ECC
    - Password cracking
- Electronic Design Automation
- Antivirus software
- Intrusion Detection

Chapter 9

Spintronics

Spintronics (a neologism meaning "spin transport electronics"), also known as magnetoelectronics, is an emerging technology that exploits both the intrinsic spin of the electron and its associated magnetic moment, in addition to its fundamental electronic charge, in solid-state devices.

History

Spintronics emerged from discoveries in the 1980s concerning spin-dependent electron transport phenomena in solid-state devices. This includes the observation of spin-polarized electron injection from a ferromagnetic metal to a normal metal by Johnson and Silsbee (1985), and the discovery of giant magnetoresistance independently by Albert Fert et al. and Peter Grünberg et al. (1988). The origins of spintronics can be traced back even further to the ferromagnet/superconductor tunneling experiments pioneered by Meservey and Tedrow, and initial experiments on magnetic tunnel junctions by Julliere in the 1970s. The use of semiconductors for spintronics can be traced back at least as far as the theoretical proposal of a spin field-effect-transistor by Datta and Das in 1990.

Theory

Electrons are spin-1/2 fermions and therefore constitute a two-state system with spin "up" and spin "down". To make a spintronic device, the primary requirements are, first, a system that can generate a current of spin-polarized electrons comprising more of one spin species—up or down—than the other (called a spin injector), and, secondly, a separate system sensitive to the spin polarization of the electrons (spin detector). Manipulation of the electron spin during transport between injector and detector (especially in semiconductors) via spin precession can be accomplished using real external magnetic fields or effective fields caused by spin-orbit interaction.

Spin polarization in non-magnetic materials can be achieved either through the Zeeman effect in large magnetic fields and low temperatures, or by non-equilibrium methods. In the latter case, the non-equilibrium polarization will decay over a timescale called the "spin lifetime". Spin lifetimes of conduction electrons in metals are relatively short (typically less than 1 nanosecond). However in semiconductors the lifetimes can be very

long (microseconds at low temperatures), especially when the electrons are isolated in local trapping potentials (for instance, at impurities, where lifetimes can be milliseconds).

Metal-based spintronic devices

The simplest method of generating a spin-polarised current in a metal is to pass the current through a ferromagnetic material. The most common application of this effect is a giant magnetoresistance (GMR) device. A typical GMR device consists of at least two layers of ferromagnetic materials separated by a spacer layer. When the two magnetization vectors of the ferromagnetic layers are aligned, the electrical resistance will be lower (so a higher current flows at constant voltage) than if the ferromagnetic layers are anti-aligned. This constitutes a magnetic field sensor.

Two variants of GMR have been applied in devices: (1) current-in-plane (CIP), where the electric current flows parallel to the layers and (2) current-perpendicular-to-plane (CPP), where the electric current flows in a direction perpendicular to the layers.

Other metals-based spintronics devices:

- Tunnel Magnetoresistance (TMR), where CPP transport is achieved by using quantum-mechanical tunneling of electrons through a thin insulator separating ferromagnetic layers.
- Spin Torque Transfer, where a current of spin-polarized electrons is used to control the magnetization direction of ferromagnetic electrodes in the device.

Applications

Motorola has developed a 1st generation 256 kb MRAM based on a single magnetic tunnel junction and a single transistor and which has a read/write cycle of under 50 nanoseconds (Everspin, Motorola's spin-off, has since developed a 4 Mbit version). There are two 2nd generation MRAM techniques currently in development: Thermal Assisted Switching (TAS) which is being developed by Crocus Technology, and Spin Torque Transfer (STT) on which Crocus, Hynix, IBM, and several other companies are working.

Another design in development, called Racetrack memory, encodes information in the direction of magnetization between domain walls of a ferromagnetic metal wire.

Semiconductor-based spintronic devices

Ferromagnetic semiconductor sources (like manganese-doped gallium arsenide GaMnAs), increase the interface resistance with a tunnel barrier, or using hot-electron injection.

Spin detection in semiconductors is another challenge, which has been met with the following techniques:

- Faraday/Kerr rotation of transmitted/reflected photons
- Circular polarization analysis of electroluminescence
- Nonlocal spin valve (adapted from Johnson and Silsbee's work with metals)
- Ballistic spin filtering

The latter technique was used to overcome the lack of spin-orbit interaction and materials issues to achieve spin transport in silicon, the most important semiconductor for electronics.

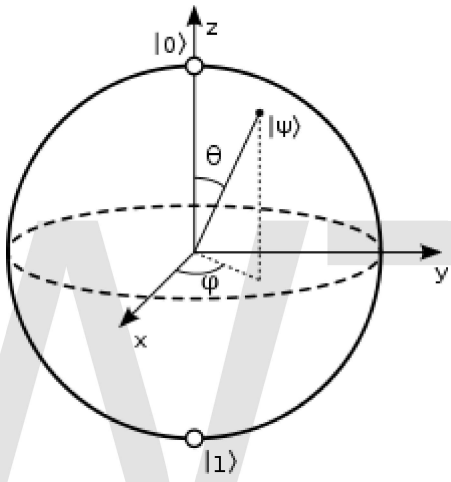
Because external magnetic fields (and stray fields from magnetic contacts) can cause large Hall effects and magnetoresistance in semiconductors (which mimic spin-valve effects), the only conclusive evidence of spin transport in semiconductors is demonstration of spin precession and dephasing in a magnetic field non-collinear to the injected spin orientation. This is called the Hanle effect.

Applications

Applications such as semiconductor lasers using spin-polarized electrical injection have shown threshold current reduction and controllable circularly polarized coherent light output. Future applications may include a spin-based transistor having advantages over MOSFET devices such as steeper sub-threshold slope.

Chapter 10

Quantum Computer



The Bloch sphere is a representation of a qubit, the fundamental building block of quantum computers.

A **quantum computer** is a device for computation that makes direct use of quantum mechanical phenomena, such as superposition and entanglement, to perform operations on data. Quantum computers are different from traditional computers based on transistors. The basic principle behind quantum computation is that quantum properties can be used to represent data and perform operations on these data. A theoretical model is the quantum Turing machine, also known as the universal quantum computer.

Although quantum computing is still in its infancy, experiments have been carried out in which quantum computational operations were executed on a very small number of qubits (quantum bits). Both practical and theoretical research continues, and many national government and military funding agencies support quantum computing research to develop quantum computers for both civilian and national security purposes, such as cryptanalysis.

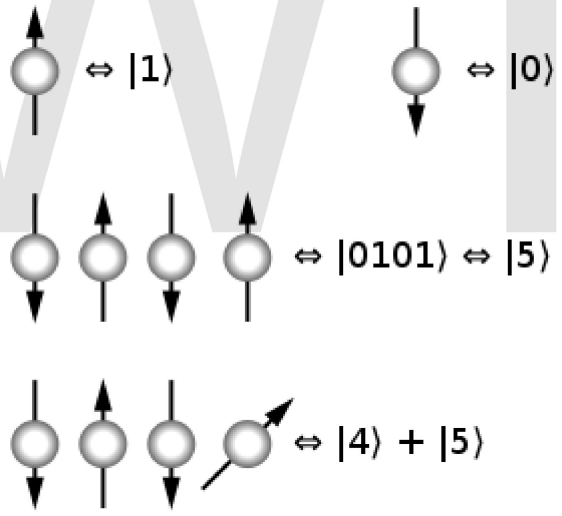
If large-scale quantum computers can be built, they will be able to solve certain problems much faster than any current classical computers (for example integer factorization using Shor's algorithm). All problems solvable with a quantum computer can also be solved using a traditional computer given enough time and resources.

Basis

A classical computer has a memory made up of bits, where each bit represents either a one or a zero. A quantum computer maintains a sequence of qubits. A single qubit can represent a one, a zero, or, crucially, any quantum superposition of these; moreover, a pair of qubits can be in any quantum superposition of 4 states, and three qubits in any superposition of 8. In general a quantum computer with *n* qubits can be in an arbitrary superposition of up to 2^n different states simultaneously (this compares to a normal computer that can only be in *one* of these 2^n states at any one time). A quantum computer operates by manipulating those qubits with a fixed sequence of quantum logic gates. The sequence of gates to be applied is called a quantum algorithm.

An example of an implementation of qubits for a quantum computer could start with the use of particles with two spin states: "down" and "up" (typically written $|\downarrow\rangle$ and $|\uparrow\rangle$, or $|0\rangle$ and $|1\rangle$). But in fact any system possessing an observable quantity *A* which is *conserved* under time evolution and such that *A* has at least two discrete and sufficiently spaced consecutive eigenvalues, is a suitable candidate for implementing a qubit. This is true because any such system can be mapped onto an effective spin-1/2 system.

Bits vs. qubits



qubits can be in a superposition of all the clasically allowed states

Qubits are made up of controlled particles and the means of control (e.g. devices that trap particles and switch them from one state to another).

Consider first a classical computer that operates on a three-bit register. The state of the computer at any time is a probability distribution over the $2^3 = 8$ different three-bit strings `000, 001, 010, 011, 100, 101, 110, 111`. If it is a deterministic computer, then it is in exactly one of these states with probability 1. However, if it is a probabilistic

computer, then there is a possibility of it being in any *one* of a number of different states. We can describe this probabilistic state by eight nonnegative numbers *a*,*b*,*c*,*d*,*e*,*f*,*g*,*h* (where *a* = probability computer is in state `000`, *b* = probability computer is in state `001`, etc.). There is a restriction that these probabilities sum to 1.

The state of a three-qubit quantum computer is similarly described by an eight-dimensional vector (*a*,*b*,*c*,*d*,*e*,*f*,*g*,*h*), called a ket. However, instead of adding to one, the sum of the *squares* of the coefficient magnitudes, $|a|^2 + |b|^2 + ... + |h|^2$, must equal one. Moreover, the coefficients are complex numbers. Since states are represented by complex wavefunctions, two states being added together will undergo interference, which is a key difference between quantum computing and probabilistic classical computing.

If you measure the three qubits, you will observe a three-bit string. The probability of measuring a given string is the squared magnitude of that string's coefficient (i.e., the probability of measuring `000` = $|a|^2$, the probability of measuring `001` = $|b|^2$, etc..). Thus, measuring a quantum state described by complex coefficients (*a*,*b*,...,*h*) gives the classical probability distribution ($|a|^2$, $|b|^2$,..., $|h|^2$) and we say that the quantum state "collapses" to a classical state as a result of making the measurement.

Note that an eight-dimensional vector can be specified in many different ways depending on what basis is chosen for the space. The basis of bit strings (e.g., 000, 001, ..., 111) is known as the computational basis. Other possible bases are unit-length, orthogonal vectors and the eigenvectors of the Pauli-x operator. Ket notation is often used to make the choice of basis explicit. For example, the state (*a*,*b*,*c*,*d*,*e*,*f*,*g*,*h*) in the computational basis can be written as:

$$a|000\rangle + b|001\rangle + c|010\rangle + d|011\rangle + e|100\rangle + f|101\rangle + g|110\rangle + h|111\rangle$$

where, e.g., $|010\rangle = (0,0,1,0,0,0,0,0)$

The computational basis for a single qubit (two dimensions) is $|0\rangle = (1,0)$ and $|1\rangle = (0,1)$.

Using the eigenvectors of the Pauli-x operator, a single qubit is $|+\rangle = \frac{1}{\sqrt{2}}(1,1)$ and $|-\rangle = \frac{1}{\sqrt{2}}(1,-1)$.

A quantum computer with a given number of qubits is exponentially more complex than a classical computer with the same number of bits because describing the state of *n* qubits requires 2^n complex coefficients. Measuring the qubits would produce a classical state of only *n* bits, but such an action would also destroy the quantum state. We can think of the system as being exactly one of the *n*-bit strings—we just don't know which one. For example, a 300-qubit quantum computer has a state described by 2^{300} (approximately 10^{90}) complex numbers, more than the number of atoms in the observable universe.

Operation

While a classical three-bit state and a quantum three-qubit state are both eight-dimensional vectors, they are manipulated quite differently for classical or quantum computation. For computing in either case, the system must be initialized, for example into the all-zeros string, $|000\rangle$, corresponding to the vector (1,0,0,0,0,0,0,0). In classical randomized computation, the system evolves according to the application of stochastic matrices, which preserve that the probabilities add up to one (i.e., preserve the L1 norm). In quantum computation, on the other hand, allowed operations are unitary matrices, which are effectively rotations (they preserve that the sum of the squares add up to one, the Euclidean or L2 norm). (Exactly what unitaries can be applied depend on the physics of the quantum device.) Consequently, since rotations can be undone by rotating backward, quantum computations are reversible. (Technically, quantum operations can be probabilistic combinations of unitaries, so quantum computation really does generalize classical computation.)

Finally, upon termination of the algorithm, the result needs to be read off. In the case of a classical computer, we *sample* from the probability distribution on the three-bit register to obtain one definite three-bit string, say 000. Quantum mechanically, we *measure* the three-qubit state, which is equivalent to collapsing the quantum state down to a classical distribution (with the coefficients in the classical state being the squared magnitudes of the coefficients for the quantum state, as described above) followed by sampling from that distribution. Note that this destroys the original quantum state. Many algorithms will only give the correct answer with a certain probability, however by repeatedly initializing, running and measuring the quantum computer, the probability of getting the correct answer can be increased.

Potential

Integer factorization is believed to be computationally infeasible with an ordinary computer for large integers if they are the product of few prime numbers (e.g., products of two 300-digit primes). By comparison, a quantum computer could efficiently solve this problem using Shor's algorithm to find its factors. This ability would allow a quantum computer to decrypt many of the cryptographic systems in use today, in the sense that there would be a polynomial time (in the number of digits of the integer) algorithm for solving the problem. In particular, most of the popular public key ciphers are based on the difficulty of factoring integers (or the related discrete logarithm problem which can also be solved by Shor's algorithm), including forms of RSA. These are used to protect secure Web pages, encrypted email, and many other types of data. Breaking these would have significant ramifications for electronic privacy and security.

However, other existing cryptographic algorithms don't appear to be broken by these algorithms. Some public-key algorithms are based on problems other than the integer factorization and discrete logarithm problems to which Shor's algorithm applies, like the McEliece cryptosystem based on a problem in coding theory. Lattice based cryptosystems are also not known to be broken by quantum computers, and finding a

polynomial time algorithm for solving the dihedral hidden subgroup problem, which would break many lattice based cryptosystems, is a well-studied open problem. It has been proven that applying Grover's algorithm to break a symmetric (secret key) algorithm by brute force requires roughly $2^{n/2}$ invocations of the underlying cryptographic algorithm, compared with roughly 2^n in the classical case, meaning that symmetric key lengths are effectively halved: AES-256 would have the same security against an attack using Grover's algorithm that AES-128 has against classical brute-force search. Quantum cryptography could potentially fulfill some of the functions of public key cryptography.

Besides factorization and discrete logarithms, quantum algorithms offering a more than polynomial speedup over the best known classical algorithm have been found for several problems, including the simulation of quantum physical processes from chemistry and solid state physics, the approximation of Jones polynomials, and solving Pell's equation. No mathematical proof has been found that shows that an equally fast classical algorithm cannot be discovered, although this is considered unlikely. For some problems, quantum computers offer a polynomial speedup. The most well-known example of this is *quantum database search*, which can be solved by Grover's algorithm using quadratically fewer queries to the database than are required by classical algorithms. In this case the advantage is provable. Several other examples of provable quantum speedups for query problems have subsequently been discovered, such as for finding collisions in two-to-one functions and evaluating NAND trees.

Consider a problem that has these four properties:

1. The only way to solve it is to guess answers repeatedly and check them,
2. The number of possible answers to check is the same as the number of inputs,
3. Every possible answer takes the same amount of time to check, and
4. There are no clues about which answers might be better: generating possibilities randomly is just as good as checking them in some special order.

An example of this is a password cracker that attempts to guess the password for an encrypted file (assuming that the password has a maximum possible length).

For problems with all four properties, the time for a quantum computer to solve this will be proportional to the square root of the number of inputs. That can be a very large speedup, reducing some problems from years to seconds. It can be used to attack symmetric ciphers such as Triple DES and AES by attempting to guess the secret key.

Grover's algorithm can also be used to obtain a quadratic speed-up over a brute-force search for a class of problems known as NP-complete.

Since chemistry and nanotechnology rely on understanding quantum systems, and such systems are impossible to simulate in an efficient manner classically, many believe quantum simulation will be one of the most important applications of quantum computing.

There are a number of practical difficulties in building a quantum computer, and thus far quantum computers have only solved trivial problems. David DiVincenzo, of IBM, listed the following requirements for a practical quantum computer:

- scalable physically to increase the number of qubits;
- qubits can be initialized to arbitrary values;
- quantum gates faster than decoherence time;
- universal gate set;
- qubits can be read easily.

Quantum decoherence

One of the greatest challenges is controlling or removing quantum decoherence. This usually means isolating the system from its environment as the slightest interaction with the external world would cause the system to decohere. This effect is irreversible, as it is non-unitary, and is usually something that should be highly controlled, if not avoided. Decoherence times for candidate systems, in particular the transverse relaxation time T_2 (for NMR and MRI technology, also called the *dephasing time*), typically range between nanoseconds and seconds at low temperature.

These issues are more difficult for optical approaches as the timescales are orders of magnitude shorter and an often-cited approach to overcoming them is optical pulse shaping. Error rates are typically proportional to the ratio of operating time to decoherence time, hence any operation must be completed much more quickly than the decoherence time.

If the error rate is small enough, it is thought to be possible to use quantum error correction, which corrects errors due to decoherence, thereby allowing the total calculation time to be longer than the decoherence time. An often cited figure for required error rate in each gate is 10^{-4}. This implies that each gate must be able to perform its task in one 10,000th of the decoherence time of the system.

Meeting this scalability condition is possible for a wide range of systems. However, the use of error correction brings with it the cost of a greatly increased number of required qubits. The number required to factor integers using Shor's algorithm is still polynomial, and thought to be between L and L^2, where L is the number of bits in the number to be factored; error correction algorithms would inflate this figure by an additional factor of L. For a 1000-bit number, this implies a need for about 10^4 qubits without error correction. With error correction, the figure would rise to about 10^7 qubits. Note that computation time is about L^2 or about 10^7 steps and on 1 MHz, about 10 seconds.

A very different approach to the stability-decoherence problem is to create a topological quantum computer with anyons, quasi-particles used as threads and relying on braid theory to form stable logic gates.

Developments

There are a number of quantum computing candidates, among those:

- Superconductor-based quantum computers (including SQUID-based quantum computers)
- Trapped ion quantum computer
- Optical lattices
- Topological quantum computer
- Quantum dot on surface (e.g. the Loss-DiVincenzo quantum computer)
- Nuclear magnetic resonance on molecules in solution (liquid NMR)
- Solid state NMR Kane quantum computers
- Electrons on helium quantum computers
- Cavity quantum electrodynamics (CQED)
- Molecular magnet
- Fullerene-based ESR quantum computer
- Optic-based quantum computers (Quantum optics)
- Diamond-based quantum computer
- Bose–Einstein condensate-based quantum computer
- Transistor-based quantum computer - string quantum computers with entrainment of positive holes using an electrostatic trap
- Spin-based quantum computer
- Adiabatic quantum computation
- Rare-earth-metal-ion-doped inorganic crystal based quantum computers

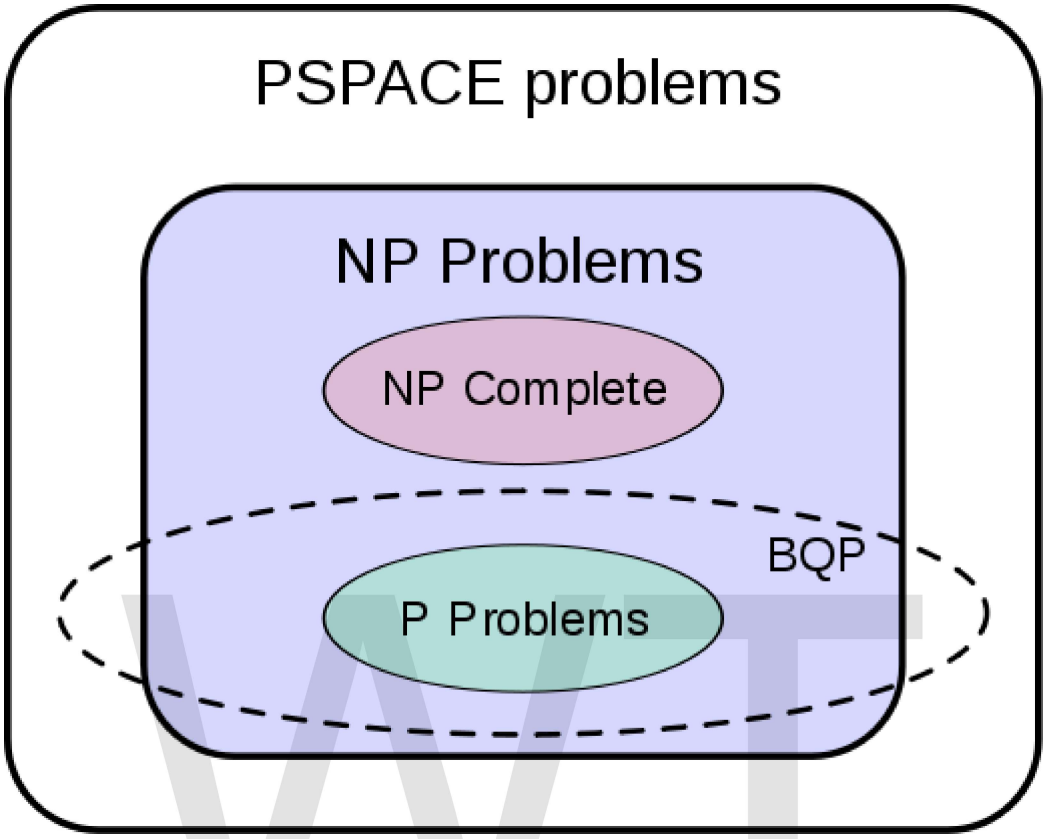
The large number of candidates demonstrates that the topic, in spite of rapid progress, is still in its infancy. But at the same time there is also a vast amount of flexibility.

In 2005, researchers at the University of Michigan built a semiconductor chip which functioned as an ion trap. Such devices, produced by standard lithography techniques, may point the way to scalable quantum computing tools. An improved version was made in 2006.

In 2009, researchers at Yale University created the first rudimentary solid-state quantum processor. The two-qubit superconducting chip was able to run elementary algorithms. Each of the two artificial atoms (or qubits) were made up of a billion aluminum atoms but they acted like a single one that could occupy two different energy states.

Another team, working at the University of Bristol, also created a silicon-based quantum computing chip, based on quantum optics. The team was able to run Shor's algorithm on the chip. The latest developments [for 2010] can be found in. Springer publish a Journal devoted to the subject.

Relation to computational complexity theory



The suspected relationship of BQP to other problem spaces

The class of problems that can be efficiently solved by quantum computers is called BQP, for "bounded error, quantum, polynomial time". Quantum computers only run probabilistic algorithms, so BQP on quantum computers is the counterpart of BPP ("bounded error, probabilistic, polynomial time") on classical computers. It is defined as the set of problems solvable with a polynomial-time algorithm, whose probability of error is bounded away from one half. A quantum computer is said to "solve" a problem if, for every instance, its answer will be right with high probability. If that solution runs in polynomial time, then that problem is in BQP.

BQP is contained in the complexity class $\#P$ (or more precisely in the associated class of decision problems $P^{\#P}$), which is a subclass of PSPACE.

BQP is suspected to be disjoint from NP-complete and a strict superset of P, but that is not known. Both integer factorization and discrete log are in BQP. Both of these problems are NP problems suspected to be outside BPP, and hence outside P. Both are suspected to not be NP-complete. There is a common misconception that quantum computers can solve NP-complete problems in polynomial time. That is not known to be true, and is generally suspected to be false.

Possibilities of the quantum computer to accelerate classical algorithms has rigid limits - lower bounds of quantum computations complexity. The overwhelming part of classical calculations can't be accelerated on the quantum computer. The similar fact takes place for particular computational tasks, like search problem, for which Grover algorithm is optimal.

Although quantum computers may be faster than classical computers, those described above can't solve any problems that classical computers can't solve, given enough time and memory (however, those amounts might be practically infeasible). A Turing machine can simulate these quantum computers, so such a quantum computer could never solve an undecidable problem like the halting problem. The existence of "standard" quantum computers does not disprove the Church–Turing thesis. It has been speculated that theories of quantum gravity, such as M-theory or loop quantum gravity, may allow even faster computers to be built. Currently, it's an open problem to even *define* computation in such theories due to the *problem of time*, i.e. there's no obvious way to describe what it means for an observer to submit input to a computer and later receive output.

Chapter 11

Technology Forecasting

Important aspects

Primarily, a technological forecast deals with the characteristics of technology, such as levels of technical performance, like speed of a military aircraft, the power in watts of a particular future engine, the accuracy or precision of a measuring instrument, the number of transistors in a chip in the year 2015, etc. The forecast does not have to state how these characteristics will be achieved.

Secondly, technological forecasting usually deals with only useful machines, procedures or techniques. This is to exclude from the domain of technological forecasting those commodities, services or techniques intended for luxury or amusement.

Rational and explicit methods

The whole purpose of the recitation of alternatives, is to show that there really is no alternative to forecasting. If a decisionmaker has several alternatives open to him, he will choose among them on the basis of which provides him with the most desirable outcome. Thus his decision is inevitably based on a forecast. His only choice is whether the forecast is obtained by rational and explicit methods, or by intuitive means.

The virtues of the use of rational methods are as follows:

1. They can be taught and learned,
2. They can be described and explained,
3. They provide a procedure followable by anyone who has absorbed the necessary training, and in some cases,
4. These methods are even guaranteed to produce the same forecast regardless of who uses them.

The virtue of the use of explicit methods is that they can be reviewed by others, and can be checked for consistency. Furthermore, the forecast can be reviewed at any subsequent time. Technology forecasting is not imagination.

Methods of technology forecasting

Commonly adopted methods of technology forecasting include the Delphi method, forecast by analogy, growth curves and extrapolation. Normative methods of technology forecasting — like the relevance trees, morphological models, and mission flow diagrams — are also commonly used.

Combining forecasts

Studies of past forecasts have shown that one of the most frequent reasons why a forecast goes wrong is that the forecaster ignores related fields.

A given technical approach may fail to achieve the level of capability forecast for it, because it is superseded by another technical approach which the forecaster ignored.

Another problem is that of inconsistency between forecasts. Because of these problems, it is often necessary to combine forecasts of different technologies. Therefore rather than to try to select the one method which is most appropriate, it may be better to try to combine the forecasts obtained by different methods.

If this is done, the strengths of one method may help compensate for the weaknesses of another.

Reasons for combining forecasts

The primary reason for combining forecasts of the same technology is to attempt to offset the weaknesses of one forecasting method with the strengths of another. In addition, the use of more than one forecasting method often gives the forecaster more insight into the processes at work which are responsible for the growth of the technology being forecast.

Trend curve and growth curves

A frequently used combination is that of growth curves and a trend curve for some technology. Here we see a succession of growth curves, each describing the level of functional capability achieved by a specific technical approach.

An overall trend curve is also shown, fitted to those items of historical data which represent the currently superior approach.

The use of growth curves and a trend curve in combination allows the forecaster to draw some conclusions about the future growth of a technology which might not be possible, were either method used alone.

With growth curves alone, the forecaster could not say anything about the time at which a given technical approach is likely to be supplanted by a successor approach.

With the trend curve alone, the forecaster could not say anything about the ability of a specific technical approach to meet the projected trend, or about the need to look for a successor approach. Thus the need for combining forecasts.

Identification of consistent deviations

Another frequently used combination of forecasts is that of the trend curve and one or more analogies.

We customarily consider the scatter of data points about a trend curve to be due to random influences which we can neither control nor even measure. However, consistent deviations may represent something other than just random influences.

Where such consistent deviations are identified, we may have an opportunity to apply an analogy. Typical events which bring about deviations from a trend are wars and depressions. Thus the purpose of combining analogies with a trend forecast is to predict deviations from the trend deviations which are associated with or caused by external events or influences.

As with other uses of analogy, it is important to determine the extent to which the analogy between the event used as the basis for the forecast, and the historical model event, satisfies the criteria for a valid analogy.

Forecasts of different technologies

Combining forecasts of different technologies may be even more important than combining the forecasts of the same technology.

One reason for this is the fact that technologies may interact or be interrelated in some fashion. Another reason for this is that of consistency in an overall picture or scenario. One of the simplest examples of interacting trends is the projection to absurdity, i.e. simply projecting the given data indefinitely without getting any specific result. For instance, if one simply projects recent rates of growth of world population, one arrives at some fantastic conclusions about the density of population in a particular place by various dates in the next millennium.

Some other trends which can confidently be expected to not continue indefinitely are:

1. Annual production of scientific papers.
2. Number of automobiles per capita.
3. Kilowatt hours of electricity generated annually.

Another instance of interacting trends was in the case of the number of scientists in the U.S. growing faster than the overall population. Since 1940s through the 1960s, science as an activity in the United States grew exponentially. The number of dollars spent on R&D was growing faster than the GNP (in the 1960s).

If projected indefinitely, these two curves would give the result that eventually every person in the U.S. would be working as a scientist and the entire GNP would be devoted to R&D alone, which are however absurd conclusions. Thus it is clear that the scientific discipline of technology forecasting is not mere trend extrapolation but also involves combining forecasts.

Uses in manufacturing

Almost all modern manufacturing firms utilize the services of a technological forecaster. Nevertheless, there are a number of alternatives to the rational and explicit forecasting of technology, such as 'no forecast', 'anything can happen' (i.e. relying on pure chance), 'window-blind forecasting', 'genius forecasting' and boasting of a 'glorious past' (i.e. adopting the same old techniques).

Thus technological forecasting is not mere astrology or palmistry, but a scientific and well defined procedure adopted by a technological forecaster or a consultancy for the forecasting of a particular technology. Even though technological forecasting is a scientific discipline, some experts are of the view that "the only certainty of a particular forecast is that it is wrong to some degree."

Leading Forecasting Institutes

- TechCast Project
- Singularity Institute for Artificial Intelligence
- Future of Humanity Institute

Chapter 12

Delphi Method

The **Delphi method** is a structured communication technique, originally developed as a systematic, interactive forecasting method which relies on a panel of experts.

In the standard version, the experts answer questionnaires in two or more rounds. After each round, a facilitator provides an anonymous summary of the experts' forecasts from the previous round as well as the reasons they provided for their judgments. Thus, experts are encouraged to revise their earlier answers in light of the replies of other members of their panel. It is believed that during this process the range of the answers will decrease and the group will converge towards the "correct" answer. Finally, the process is stopped after a pre-defined stop criterion (e.g. number of rounds, achievement of consensus, stability of results) and the mean or median scores of the final rounds determine the results.

Other versions, such as the Policy Delphi, have been designed for normative and explorative use, particularly in the area of social policy and public health. In Europe, more recent web-based experiments have used the Delphi method as a communication technique for interactive decision-making and e-democracy.

Delphi is based on the principle that forecasts (or decisions) from a structured group of individuals are more accurate than those from unstructured groups. This has been indicated with the term "collective intelligence". The technique can also be adapted for use in face-to-face meetings, and is then called mini-Delphi or Estimate-Talk-Estimate (ETE). Delphi has been widely used for business forecasting and has certain advantages over another structured forecasting approach, prediction markets.

History

The name "Delphi" derives from the Oracle of Delphi. The authors of the method were not happy with this name, because it implies "something oracular, something smacking a little of the occult". The Delphi method is based on the assumption that group judgments are more valid than individual judgments.

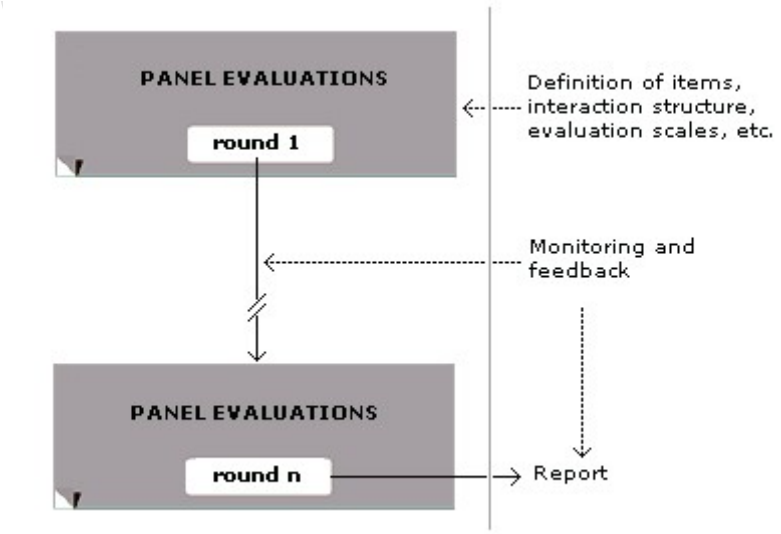
The Delphi method was developed at the beginning of the Cold War to forecast the impact of technology on warfare. In 1944, General Henry H. Arnold ordered the creation

of the report for the U.S. Army Air Corps on the future technological capabilities that might be used by the military.

Different approaches were tried, but the shortcomings of traditional forecasting methods, such as theoretical approach, quantitative models or trend extrapolation, in areas where precise scientific laws have not been established yet, quickly became apparent. To combat these shortcomings, the Delphi method was developed by Project RAND during the 1950-1960s (1959) by Olaf Helmer, Norman Dalkey, and Nicholas Rescher. It has been used ever since, together with various modifications and reformulations, such as the Imen-Delphi procedure.

Experts were asked to give their opinion on the probability, frequency, and intensity of possible enemy attacks. Other experts could anonymously give feedback. This process was repeated several times until a consensus emerged.

Key characteristics



The Delphi Method communication structure

The following key characteristics of the Delphi method help the participants to focus on the issues at hand and separate Delphi from other methodologies:

Structuring of information flow

The initial contributions from the experts are collected in the form of answers to questionnaires and their comments to these answers. The panel director controls the interactions among the participants by processing the information and filtering out irrelevant content. This avoids the negative effects of face-to-face panel discussions and solves the usual problems of group dynamics.

Regular feedback

Participants comment on their own forecasts, the responses of others and on the progress of the panel as a whole. At any moment they can revise their earlier statements. While in regular group meetings participants tend to stick to previously stated opinions and often conform too much to group leader, the Delphi method prevents it.

Anonymity of the participants

Usually all participants remain anonymous. Their identity is not revealed, even after the completion of the final report. This prevents the authority, personality, or reputation of some participants from dominating others in the process. Arguably, it also frees participants (to some extent) from their personal biases, minimizes the "bandwagon effect" or "halo effect", allows free expression of opinions, encourages open critique, and facilitates admission of errors when revising earlier judgments.

Role of the facilitator

The person coordinating the Delphi method can be known as a *facilitator*, and facilitates the responses of their *panel of experts*, who are selected for a reason, usually that they hold knowledge on an opinion or view. The facilitator sends out questionnaires, surveys etc. and if the panel of experts accept, they follow instructions and present their views. Responses are collected and analyzed, then common and conflicting viewpoints are identified. If consensus is not reached, the process continues through thesis and antithesis, to gradually work towards synthesis, and building consensus.

Use in forecasting

First applications of the Delphi method were in the field of science and technology forecasting. The objective of the method was to combine expert opinions on likelihood and expected development time, of the particular technology, in a single indicator. One of the first such reports, prepared in 1964 by Gordon and Helmer, assessed the direction of long-term trends in science and technology development, covering such topics as scientific breakthroughs, population control, automation, space progress, war prevention and weapon systems. Other forecasts of technology were dealing with vehicle-highway systems, industrial robots, intelligent internet, broadband connections, and technology in education.

Later the Delphi method was applied in other areas, especially those related to public policy issues, such as economic trends, health and education. It was also applied successfully and with high accuracy in business forecasting. For example, in one case reported by Basu and Schroeder (1977), the Delphi method predicted the sales of a new product during the first two years with inaccuracy of 3–4% compared with actual sales. Quantitative methods produced errors of 10–15%, and traditional unstructured forecast methods had errors of about 20%.

The Delphi method has also been used as a tool to implement multi-stakeholder approaches for participative policy-making in developing countries. The governments of Latin America and the Caribbean have successfully used the Delphi method as an open-ended public-private sector approach to identify the most urgent challenges for their regional ICT-for-development eLAC Action Plans. As a result, governments have widely acknowledged the value of collective intelligence from civil society, academic and private sector participants of the Delphi, especially in a field of rapid change, such as technology policies. In this sense, the Delphi method can contribute to a general appreciation of participative policy-making.

Acceptance

Overall the track record of the Delphi method is mixed. There have been many cases when the method produced poor results. Still, some authors attribute this to poor application of the method and not to the weaknesses of the method itself. It must also be realized that in areas such as science and technology forecasting the degree of uncertainty is so great that exact and always correct predictions are impossible, so a high degree of error is to be expected.

Another particular weakness of the Delphi method is that future developments are not always predicted correctly by consensus of experts. Firstly, the issue of ignorance is important. If panelists are misinformed about a topic, the use of Delphi may only add confidence to their ignorance. Secondly, sometimes unconventional thinking of amateur outsiders may be superior to expert thinking.

One of the initial problems of the method was its inability to make complex forecasts with multiple factors. Potential future outcomes were usually considered as if they had no effect on each other. Later on, several extensions to the Delphi method were developed to address this problem, such as cross impact analysis, that takes into consideration the possibility that the occurrence of one event may change probabilities of other events covered in the survey. Still the Delphi method can be used most successfully in forecasting single scalar indicators.

Despite these shortcomings, today the Delphi method is a widely accepted forecasting tool and has been used successfully for thousands of studies in areas varying from technology forecasting to drug abuse.

Use in policy-making

From the 1970's, the use of the Delphi technique in public policy-making introduces a number of methodological innovations. In particular:

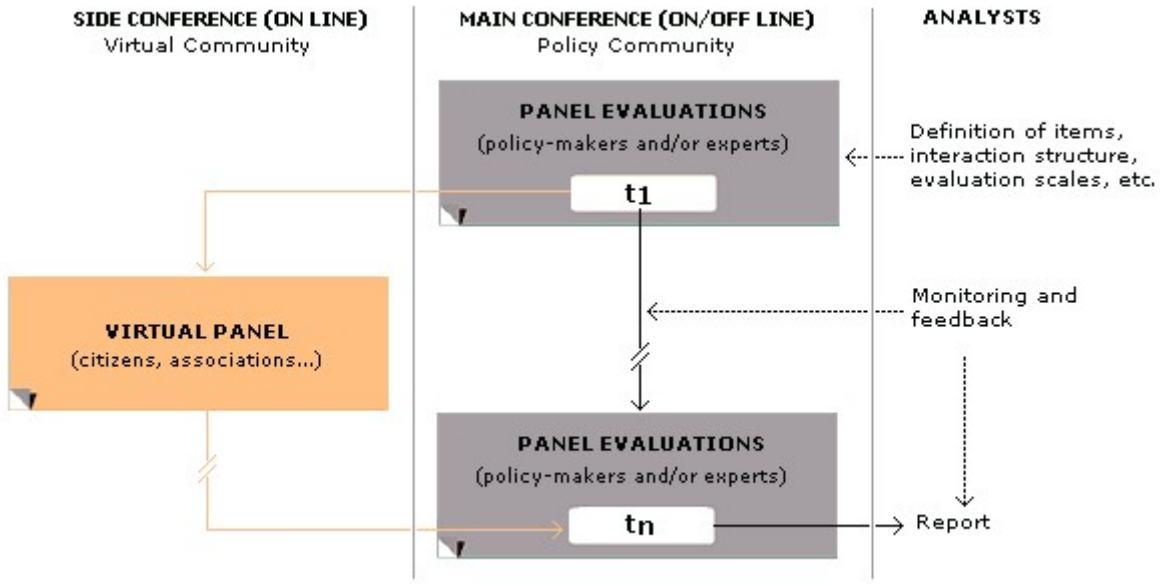
- the need to examine several types of items (not only *forecasting* items but, typically, *issue* items, *goal* items, and *option* items) leads to introducing different evaluation scales which are not used in the standard Delphi. These often include *desirability*, *feasibility* (technical and political) and *probability*, which the

analysts can use to outline different scenarios: the *desired* scenario (from desirability), the *potential* scenario (from feasibility) and the *expected* scenario (from probability);
- the complexity of the issues posed in public policy-making leads to give more importance to the arguments supporting the evaluations of the panelists; so these are often invited to list arguments for and against each option item, and sometimes they are given the possibility to suggest new items to be submitted to the panel;
- for the same reason, the scaling methods, which are used to measure panel evaluations, often include more sophisticated approaches such as multi-dimensional scaling.

Further innovations come from the use of computer-based (and later web-based) Delphi conferences. According to Turoff and Hiltz, in computer-based Delphis:

- the iteration structure used in the paper Delphis, which is divided into three or more discrete rounds, can be replaced by a process of continuous (roundless) interaction, enabling panelists to change their evaluations at any time;
- the statistical group response can be updated in real-time, and shown whenever a panelist provides a new evaluation.

According to Bolognini, web-based Delphis offer two further possibilities, relevant in the context of interactive policy-making and e-democracy. These are:



A web-based communication structure (Hyperdelphi)

- the involvement of a large number of participants,

- the use of two or more panels representing different groups (such as policy-makers, experts, citizens), which the administrator can give tasks reflecting their diverse roles and expertise, and make them to interact within ad hoc communication structures. For example, the *policy community* members (policy-makers and experts) may interact as part of the *main conference* panel, while they receive inputs from a *virtual community* (citizens, associations etc.) involved in a *side conference*. These web-based variable communication structures, which he calls *Hyperdelphi* (HD), are designed to make Delphi conferences "more fluid and adapted to the hypertextual and interactive nature of digital communication".

Delphi applications not aiming at consensus

Traditionally the Delphi method has aimed at a consensus of the most probable future by iteration. The Policy Delphi, launched by Murray Turoff, is instead a decision support method aiming at structuring and discussing the diverse views of the preferred future. The Argument Delphi, developed by Osmo Kuusi, focuses on ongoing discussion and finding relevant arguments rather than focusing on the output. The Disaggregative Policy Delphi, developed by Petri Tapio, uses cluster analysis as a systematic tool to construct various scenarios of the future in the latest Delphi round. The respondent's view on the probable and the preferable future are dealt with as separate cases.

Delphi vs. prediction markets

As can be seen from the Methodology Tree of Forecasting, Delphi has characteristics similar to prediction markets as both are structured approaches that aggregate diverse opinions from groups. Yet, there are differences that may be decisive for their relative applicability for different problems.

Some advantages of prediction markets derive from the possibility to provide incentives for participation.

1. They can motivate people to participate over a long period of time and to reveal their true beliefs.
2. They aggregate information automatically and instantly incorporate new information in the forecast.
3. Participants do not have to be selected and recruited manually by a facilitator. They themselves decide whether to participate if they think their private information is not yet incorporated in the forecast.

Delphi seems to have these advantages over prediction markets:

1. Potentially quicker forecasts if experts are readily available.

Online Delphi forecasting systems

A number of Delphi forecasts are conducted using web sites that allow the process to be conducted in Real-time Delphi. For instance, the TechCast Project uses a panel of 100 experts worldwide to forecast breakthroughs in all fields of science and technology. Further examples are several studies conducted by the Center for Futures Studies and Knowledge Management that use an online-based Delphi method.

Software

- *Principles of Forecasting* A free service to support Delphi forecasting and references are available on this site. However source code is not currently available.
- *Center for Futures Studies and Knowledgemanagement* Access to several free studies that illustrate the use of the Real-time Delphi method

Chapter 13

Real-time Delphi

Real-time Delphi (RTD) is an advanced form of the Delphi method. The advanced method "is a consultative process that uses computer technology" to increase efficiency of the Delphi process.

Definition and Idea

Gordon and Pease define the advanced approach as an innovative way to conduct Delphi studies that do not involve sequential "rounds" and consequently lead to a higher degree of efficiency with regard to the time frame needed to perform such studies. Friedewald, von Oertzen and Cuhls underline that aspect by writing, in "a Real-Time-Delphi, the participants do not only judge twice but can change their opinion as often as they like when they see the aggregated results of the other participants". So, here it becomes clear that the Real-Time Delphi approach requires real-time calculation and provision of group responses. Friedewald et al. further state that the Real-Time Delphi method has beneath its explorative and predictive elements also normative and communicative elements. These latter are investigated by Bolognini, who explores the potential of computer-based Delphi as a communication technique for electronic democracy.

History

The basic idea of a real-time, therefore computer-based (usually web-based), Delphi approach originates in a paper published by Turoff back in 1972 about an online Delphi conference conducted in the United States. The conference was characterized by remote locations of participants, an online tool to access and give judgments, anonymity of the participants, continuous operations and analysis of results (i.e. participants were able to see given answers of the other participants in real-time), as well as asynchronous participation (i.e. participants could independently login and logout how often and when they desired). The stated aspects are some of the key characteristics of Real-Time Delphi studies, which shows that the original idea of conducting such studies can be traced back to the respective year. Today, nevertheless, technological innovations and advanced computer aided design possibilities (e.g. high-speed internet connections, high definition graphic, and advanced processor performance) facilitate more sophisticated studies in this context. The general idea to develop a faster advanced form of Delphi studies by using ideas and basic concepts of Turoff, was recently initiated by the U.S. Defense Advanced Research Projects Agency (DARPA), which awarded a grant in 2004 to develop an

approach to improve “speed and efficiency of collecting judgments in tactical situations”. A small software company named Articulate Software in San Francisco was awarded an innovation research grant to develop what DARPA was asking for. Adam Pease, principal consultant and CEO of Articulate Software, published the findings and methodology together with Theodore Gordon in 2006.

Differences between Conventional and Real-time Delphi Method

The basic framework is to think of a Delphi study which is conducted in form of an online questionnaire. However, a Conventional round-based Delphi study conducted via the internet is called “Internet Delphi”. The basic difference to Internet Delphi is that the process of a Real-Time Delphi is not characterised by single iterated rounds. In fact, real-time calculation and provision of responses are the key characteristics of Real-Time Delphis. Various other labels for Real-Time Delphi can be found in literature and many authors are not completely aware of the differences: “Electronic Delphi”, “Computer Delphi”, “Computer-aided Delphi” and “Technology Delphi”. However, it is important to truly understand the design and process a researcher has chosen to find out whether real-time calculations and provisions have been applied or not.

The typical Real-Time Delphi process can be described in the way that participants get access to an online questionnaire portal for a certain time frame, within which they are allowed to login and logout as often as they want. Whenever they login, they will see all their quantitative and qualitative answers of previous sessions and they can change all answers as desired within the given period of time. Besides their own answers they will see the on-going – hence, real-time – responses of other participants, and with regard to metric assessments the group as a whole will be visualised in terms of median, average, and interquartile range (IQR). It has to be pointed out that the numerical visualisations as well as the qualitative inputs change in the course of other participants changing their responses. Consequently, a participant can find out to what extent his own responses from an earlier point of time are still within the group opinion (i.e. IQR). The core innovation, then, of Real-Time Delphi studies is the real-time calculation and provision of results.

Methodological Advancements

The core methodological innovation of Real-Time Delphi studies are the absence of iterated rounds and the real-time calculation and provision of group responses. Whereas Conventional Delphi studies are characterised by repeating sequential rounds, the Real-Time Delphi approach is characterised by a continuous round-less procedure leading to a reduced time frame needed to conduct such studies. Consequently, conducting large-scale studies of huge complexity in a relatively short period of time becomes possible. Another core methodological innovation is the fact that experts may not only judge once or twice, depending on the number of rounds, as it has been usual in a Conventional Delphi study. During a Real-Time Delphi, experts can independently reassess their responses as often as they want.

Hartman and Baldwin discuss further advantages of the Real-Time Delphi approach: First, the number of experts participating in the real-time study can be increased due to a higher degree of automation during and improved possibilities for analysis after the study. Additionally, the Internet provides the possibility to invite a worldwide expert panel to participate in the study. Second, the degree of interaction among the experts can be increased due to the fact that they can immediately react on others' comments. Additionally, the time frame between giving own answers and getting insights into others' responses is very short, which encourages stronger cognitive examination with the respective issue in question. Hartman and Baldwin argue that with the help of this procedure the validity of results is maximised.

In order to conduct a Real-Time Delphi study, computer software – usually web-based – is needed for facilitating real-time calculations and visualisation of results. It is generally proposed in the existing literature that the experts participating in the study see not only their own answers but also the median and interquartile range of all given responses immediately after answering a quantitative question. Besides the quantitative assessment a qualitative judgment of participants can be shown which serves as a justification for their numerical assessment of the question. Additionally, it can be shown to the expert how many respondents have already given their answers. To examine the qualitative arguments of others participants can click on a button, and a "reasons window" opens, which shows the statements of other participants to underline their point of view. So, the respective legitimisations given by others may cause a respondent to recapture his own point of view. In the next step the expert can change his own answer, add new arguments to underline his point of view, or leave his answer unchanged. In addition, the respondent will be shown an attention indicator, a so called "flag", if his answer is within or without the interquartile range or significantly different from the median. This application helps to see and understand immediately the own assessment and to think about reasons for the deviation from group opinions or else a high degree of consent. The respondent's attention will be called by highlighting questions with a high degree of deviation with a different colour and by asking him to give further reasons for his deviation from the group opinion. After operating a question in the described procedure, the participant can continue to the next question or press a "save"-button in the program, which leads to an immediate update of the median, interquartile range, and given arguments, and then leave the program. A second advantage of the round-less approach is the fact that, in order to take part in the study, participants can login and logout with their personalised account as often as they want during the time frame provided. Their already given answers will be saved and recalled when they login the next time. So, by design of the study, there are no explicit single rounds to answer the questions. Updating and playing back the information to the other participants follow immediately in succession to the process of answering. Here, it becomes clear that the process of answering can be synchronous or asynchronous and a worldwide expert panel can be reached, which is one of the major advantages of web-based tools. Turoff and Hiltz argue that the issue of asynchronous interaction is probably one of the least understood characteristics of Real-Time Delphis. Zipfinger points out two advantages of asynchronous participation of the experts: First, they can login to the portal whenever they want; therefore, one could argue that the degree of convenience of taking part is increased for the participant due to a 24h-availability of the

portal. Second, panellists can contribute to whatever aspects in the questionnaire they want, especially when having gone through each question at least once. Here, a substantial aspect of Real-Time Delphis becomes obvious: Turoff and Hiltz explain that a Real-Time Delphi study offers a design of structured communication which allows every individual to choose the sequence and speed to contribute to the problem solution process. So, in comparison to face-to-face discussions, the Real-Time Delphi approach gives room for individuality and different cognitive abilities of the participants. A further advantage is the fact that the administrator of the study can set an arbitrary time frame in which participants have to login and take part in the questionnaire. So, whenever the researcher or administrator of the study is satisfied with the existing answers (i.e. in terms of quantity and quality), he can declare the study to be ended and close the online tool (i.e. "freeze" the responses). It is obvious that the key features of Delphi studies are also met in the context of Real-Time Delphi studies. However, the issue of iteration is, by design of the technique, not valid for Real-Time Delphi studies anymore. Instead of answering each question a first time and getting a second sheet with the group responses in the second round, the Real-Time Delphi already shows the second screen (i.e. group responses) immediately after answering each question. Having answered each question at least once, the participant can usually control which question to reassess from a "consensus portal", which serves as a kind of control panel to access single questions again. So, on the one hand, the procedure differs from a Conventional Delphi and, on the other hand, the iteration into single rounds is missing. Having asked the question how the accomplishment of a Real-Time Delphi study differs from conducting a usual Delphi study, Gordon and Pease point out that a Real-Time Delphi study can be implemented via a site on the Internet or in any other network (e.g. intra-company network, local area network) and is, therefore, not conducted in paper-and-pencil form any more. As with all Delphi studies, the process of defining and selecting experts is still extremely important. The Conventional Delphi study is then divided into several steps of response round, analysis through the facilitator, playing back the information, next response round, and so on. However, the Real-Time Delphi study is, after granting access to the online tool, rather a self-running process. The basic strengths of a Real-Time Delphi study are its efficiency and applicability to all Delphi topics (i.e. common problem sets, decision making issues, cross impact studies, etc.). Figure 2 illustrates that the process of a Real-Time Delphi differs. Important is to point out that the number of interventions of the facilitator needed during the response phases (i.e. after opening the online tool) are usually less. Having developed the online tool in advance, the intermediate analysis done by the facilitator of the study is rather uncomplicated in comparison to the Conventional Delphi. The overall shortened time period needed to conduct a real-time study underlines that the approach can be regarded as generally more efficient. Gordon and Pease point out that a Real-Time Delphi study is applicable for a wide range of possible circumstances under which the consultation of experts is necessary. On the one hand, the authors give the example of a "small group operating synchronously in a conference room with laptop computers connected wirelessly to the web site where the software resides, with anticipated completion of the exercise in say 20 min." On the other hand, it can be thought of a larger panel of experts operating asynchronously from remote locations within a longer period of time. The greatest weakness of the real-time approach is that it is missing a wholly integrated, scientifically founded concept. The real-time

Delphi idea is still a very new concept, which requires further research and application to become a tool for full scale operations. Especially the editing of the alpha (i.e. the first) inputs of respondents, the real-time presentation of group results, and the tracking of progress over time should be integrated in a kind of administrator package to make the accomplishment of a Real-Time Delphi less difficult.

Chapter 14

Virtual Reality

U.S. Navy personnel using a VR parachute trainer

World Skin (1997), Maurice Benayoun's virtual reality interactive installation

Virtual reality (**VR**) is a term that applies to computer-simulated environments that can simulate physical presence in places in the real world, as well as in imaginary worlds. Most current virtual reality environments are primarily visual experiences, displayed either on a computer screen or through special stereoscopic displays, but some simulations include additional sensory information, such as sound through speakers or headphones. Some advanced, haptic systems now include tactile information, generally known as force feedback, in medical and gaming applications. Furthermore, virtual reality covers remote communication environments which provide virtual presence of users with the concepts of telepresence and telexistence.

Users can interact with a virtual environment or a virtual artifact (VA) either through the use of standard input devices such as a keyboard and mouse, or through multimodal devices such as a wired glove, the Polhemus, and omnidirectional treadmills. The simulated environment can be similar to the real world—for example, in simulations for pilot or combat training—or it can differ significantly from reality, such as in VR games. In practice, it is currently very difficult to create a high-fidelity virtual reality experience, due largely to technical limitations on processing power, image resolution, and communication bandwidth; however, the technology's proponents hope that such limitations will be overcome as processor, imaging, and data communication technologies become more powerful and cost-effective over time.

Virtual reality is often used to describe a wide variety of applications commonly associated with immersive, highly visual, 3D environments. The development of CAD software, graphics hardware acceleration, head mounted displays, database gloves, and miniaturization have helped popularize the notion. In the book *The Metaphysics of Virtual Reality* by Michael R. Heim, seven different concepts of virtual reality are identified: simulation, interaction, artificiality, immersion, telepresence, full-body immersion, and network communication. People often identify VR with head mounted displays and data suits.

Background

Terminology and concepts

The term "artificial reality", coined by Myron Krueger, has been in use since the 1970s; however, the origin of the term "virtual reality" can be traced back to the French playwright, poet, actor, and director Antonin Artaud. In his seminal book *The Theatre and Its Double* (1938), Artaud described theatre as "*la réalite virtuelle*", a virtual reality "in which characters, objects, and images take on the phantasmagoric force of alchemy's visionary internal dramas". It has been used in *The Judas Mandala*, a 1982 science-fiction novel by Damien Broderick, where the context of use is somewhat different from that defined above. The earliest use cited by the Oxford English Dictionary is in a 1987 article titled "Virtual reality", but the article is not about VR technology. The concept of virtual reality was popularized in mass media by movies such as *Brainstorm* and *The Lawnmower Man*. The VR research boom of the 1990s was accompanied by the non-fiction book *Virtual Reality* (1991) by Howard Rheingold. The book served to demystify the subject, making it more accessible to less technical researchers and enthusiasts, with an impact similar to that which his book *The Virtual Community* had on virtual community research lines closely related to VR. *Multimedia: from Wagner to Virtual Reality*, edited by Randall Packer and Ken Jordan and first published in 2001, explores the term and its history from an avant-garde perspective. Philosophical implications of the concept of VR are systematically discussed in the book *Get Real: A Philosophical Adventure in Virtual Reality* (1998) by Philip Zhai, wherein the idea of VR is pushed to its logical extreme and ultimate possibility. According to Zhai, virtual reality could be made to have an ontological status equal to that of actual reality. *Digital Sensations: Space, Identity and Embodiment in Virtual Reality* (1999), written by Ken Hillis, offers a more critical and theoretical academic assessment of the complex set of cultural and political desires and practices culminating in the development of the technology.

Timeline

Virtual reality can trace its roots to the 1860s, when 360-degree art through panoramic murals began to appear. An example of this would be Baldassare Peruzzi's piece titled, *Sala delle Prospettive*. In the 1920s, vehicle simulators were introduced. Morton Heilig wrote in the 1950s of an "Experience Theatre" that could encompass all the senses in an effective manner, thus drawing the viewer into the onscreen activity. He built a prototype of his vision dubbed the Sensorama in 1962, along with five short films to be displayed in

it while engaging multiple senses (sight, sound, smell, and touch). Predating digital computing, the Sensorama was a mechanical device, which reportedly still functions today. Around this time, Douglas Englebart uses computer screens as both input and output devices. In 1966, Tom Furness introduces a visual flight stimulator for the Air Force. In 1968, Ivan Sutherland, with the help of his student Bob Sproull, created what is widely considered to be the first virtual reality and augmented reality (AR) head mounted display (HMD) system. It was primitive both in terms of user interface and realism, and the HMD to be worn by the user was so heavy it had to be suspended from the ceiling. The graphics comprising the virtual environment were simple wireframe model rooms. The formidable appearance of the device inspired its name, The Sword of Damocles. Also notable among the earlier hypermedia and virtual reality systems was the Aspen Movie Map, which was created at MIT in 1977. The program was a crude virtual simulation of Aspen, Colorado in which users could wander the streets in one of three modes: summer, winter, and polygons. The first two were based on photographs—the researchers actually photographed every possible movement through the city's street grid in both seasons—and the third was a basic 3-D model of the city. In the late 1980s, the term "virtual reality" was popularized by Jaron Lanier, one of the modern pioneers of the field. Lanier had founded the company VPL Research in 1985, which developed and built some of the seminal "goggles and gloves" systems of that decade. In 1991, Antonio Medina, a MIT graduate and NASA scientist, designed a virtual reality system to "drive" Mars rovers from Earth in apparent real time despite the substantial delay of Mars-Earth-Mars signals. The system, termed "Computer-Simulated Teleoperation" as published by Rand, is an extension of virtual reality.

Impact

There has been an increase in interest in the potential social impact of new technologies, such as virtual reality. Mychilo S. Cline, in his book *Power, Madness, and Immortality: The Future of Virtual Reality*, argues that virtual reality will lead to a number of important changes in human life and activity. He argues that:

- Virtual reality will be integrated into daily life and activity, and will be used in various human ways.
- Techniques will be developed to influence human behavior, interpersonal communication, and cognition.
- As we spend more and more time in virtual space, there will be a gradual "migration to virtual space", resulting in important changes in economics, worldview, and culture.
- The design of virtual environments may be used to extend basic human rights into virtual space, to promote human freedom and well-being, and to promote social stability as we move from one stage in socio-political development to the next.
- Virtual reality can also be used to induce body transfer illusions.

Heritage and archaeology

The use of VR in heritage and archaeology has potential in museum and visitor centre applications, but its use has been tempered by the difficulty in presenting a "quick to learn" real time experience to numerous people at any given time. Many historic reconstructions tend to be in a pre-rendered format to a shared video display, thus allowing more than one person to view a computer generated world, but limiting the interaction that full-scale VR can provide. The first use of a VR presentation in a heritage application was in 1994, when a museum visitor interpretation provided an interactive "walk-through" of a 3D reconstruction of Dudley Castle in England as it was in 1550. This consisted of a computer controlled laserdisc-based system designed by British based engineer Colin Johnson. One of the first users of virtual reality was Queen Elizabeth II, when she officially opened this visitor centre in June 1994. The system was featured in a conference held by the British Museum in November 1994, and in the subsequent technical paper, *Imaging the Past - Electronic Imaging and Computer Graphics in Museums and Archaeology*.

VR reconstruction

Virtual reality enables heritage sites to be recreated extremely accurately, so that the recreations can be published in various media. The original sites are often inaccessible to the public, or may even no longer exist. This technology can be used to develop virtual replicas of caves, natural environment, old towns, monuments, sculptures and archaeological elements.

Fiction books

Many science fiction books and movies have imagined characters being "trapped in virtual reality".

A comprehensive and specific fictional model for virtual reality was published in 1935 in the short story Pygmalion's Spectacles by Stanley G. Weinbaum. In the story, the main character, Dan Burke, meets an elfin professor, Albert Ludwig, who has invented a pair of goggles which enable "a movie that gives one sight and sound [...] taste, smell, and touch. [...] You are in the story, you speak to the shadows (characters) and they reply, and instead of being on a screen, the story is all about you, and you are in it." A more modern work to use this idea was Daniel F. Galouye's novel Simulacron-3, which was made into a German teleplay titled *Welt am Draht* ("World on a Wire") in 1973. Other science fiction books have promoted the idea of virtual reality as a partial, but not total, substitution for the misery of reality, or have touted it as a method for creating breathtaking virtual worlds in which one may escape from Earth.

Stanisław Lem wrote a short story in early 1960 called "*dziwne skrzynie profesora Corcorana*", in which he presented a scientist who devised a completely artificial virtual reality. Among the beings trapped inside his created virtual world, there is also a scientist, who also devised such machines creating another level of virtual world. The

Piers Anthony novel *Killobyte* follows the story of a paralyzed cop trapped in a virtual reality game by a hacker, whom he must stop to save a fellow trapped player slowly succumbing to insulin shock. This novel toys with the idea of both the potential positive therapeutic uses, such as allowing the paralyzed to experience the illusion of movement while stimulating unused muscles, as well as virtual realities' dangers. Vernor Vinge's *True Names*, published in 1981, imagines a virtual world which is probably the first to represent a metaverse. In the story, characters interact with each other in a complete world, where they own homes and are represented using avatars. This type of virtual world was later to be realized as Second Life, which was launched in 2003.

Other popular fictional works that use the concept of virtual reality include William Gibson's *Neuromancer* which defined the concept of cyberspace, Neal Stephenson's *Snow Crash*, in which he made extensive reference to the term avatar to describe one's representation in a virtual world, and Rudy Rucker's *The Hacker and the Ants*, in which programmer Jerzy Rugby uses VR for robot design and testing.

The *Doctor Who* serial "The Deadly Assassin", first broadcast in 1976, introduced a dream-like computer-generated reality, known as the Matrix.

The first major American television series to showcase virtual reality regularly was *Star Trek: The Next Generation*. Several episodes featured a holodeck, a virtual reality facility that enabled its users to recreate and experience anything they wanted. One difference from current virtual reality technology, however, was that replicators, force fields, holograms, and transporters were used to actually recreate and place objects in the holodeck, rather than illusions.

The New Zealand post-apocalyptic soap opera The Tribe (TV Series) shows Virtual Reality being used by an advanced enemy tribe named the Technos.

Cult British BBC2 sci-fi series Red Dwarf featured a virtual reality game titled "Better Than Life", in which the main characters had spent many years connected. Virtual reality has also been featured in other Red Dwarf episodes, including "Back to Reality", where venom from the despair squid caused the characters to believe that all of their experiences on Red Dwarf had been part of a VR simulation. Other episodes that feature virtual reality include "Gunmen of the Apocalypse", "Stoke Me a Clipper", "Blue", "Beyond a Joke", and "Back in the Red".

The popular .hack multimedia franchise is based on a virtual reality MMORPG dubbed "The World" The French animated series *Code Lyoko* is based on the virtual world of *Lyoko* and the Internet. The virtual world is accessed by large scanners which use an atomic process, and breaks down the atoms of the person inside, digitizes them, and recreates an incarnation on *Lyoko*. And there was also the Saban show VR Troopers.

Motion pictures

Steven Lisberger's 1982 movie *TRON* was the first mainstream Hollywood picture to explore the idea of virtual reality. One year later, it would be fully expanded in the Natalie Wood film *Brainstorm*. A VR-like system used to record and play back dreams figures centrally in Wim Wenders' 1991 film *Until the End of the World*. The 1992 film *The Lawnmower Man* (which bore almost no resemblance to the Stephen King story of the same name on which it was ostensibly based) tells the story of a research scientist who uses a VR system to jumpstart the mental and physical development of his mentally handicapped gardener. James Cameron's *Avatar* depicts a time when people's consciousness are virtually transported into biologically grown avatars. Outside the genre of science fiction, 1994's *Disclosure*, starring Michael Douglas, based on the Michael Crichton book of the same name, depicts a VR headset being used as a navigation device for a prototype computer file system.

Games

Classic Virtual reality HMD with glove

In 1991, Virtuality (originally W Industries) licensed the Amiga 3000 for use in their VR machines, and released a VR gaming system called the 1000CS. This was a stand-up immersive HMD platform with a tracked 3D joystick. The system featured several VR games including *Dactyl Nightmare*, *Legend Quest*, *Hero*, and *Grid Busters*. The Aura Interactor Virtual Reality Game Wear is a chest and back harness through which the player can feel punches, explosions, kicks, uppercuts, slam-dunks, crashes, and bodyblows. It works with the Sega Genesis and Super Nintendo Entertainment System.

In the *Mage: The Ascension* role-playing game, the mage tradition of the Virtual Adepts is presented as the creators of VR. The Adepts' ultimate objective is to move into virtual reality, scrapping their physical bodies in favour of improved virtual ones. Also, the *.hack* series centers on a virtual reality video game. This shows the potentially dangerous side of virtual reality, demonstrating the adverse effects on human health and possible viruses, including a comatose state which some players assume. Metal Gear Solid bases heavily on VR usage, either as a part of the plot (notably Metal Gear Solid 2), or simply to guide the players through training sessions.

Radio

In 2009, British digital radio station BBC Radio 7 broadcasted *Planet B*, a science-fiction drama set in a virtual world. *Planet B* is the largest ever commission for an original drama programme.

Fine art

David Em was the first fine artist to create navigable virtual worlds in the 1970s. His early work was done on mainframes at III, JPL, and Caltech. Jeffrey Shaw explored the potential of VR in fine arts with early works like *Legible City* (1989), *Virtual Museum* (1991), and *Golden Calf* (1994). Canadian artist Char Davies created immersive VR art pieces *Osmose* (1995) and *Ephémère* (1998). Maurice Benayoun's work introduced metaphorical, philosophical or political content, combining VR, network, generation and intelligent agents, in works like *Is God Flat* (1994), *The Tunnel under the Atlantic* (1995), and *World Skin* (1997). Other pioneering artists working in VR have include Luc Courchesne, Rita Addison, Knowbotic Research, Rebecca Allen, Perry Hoberman, Jacki Morie, and Brenda Laurel. All mentioned artists are documented in the Database of Virtual Art.

Therapeutic uses

The primary use of VR in a therapeutic role is its application to various forms of exposure therapy, ranging from phobia treatments to newer approaches to treating PTSD. A very basic VR simulation with simple sight and sound models has been shown to be invaluable in phobia treatment, like zoophobia, and acrophobia, as a step between basic exposure therapy such as the use of simulacra and true exposure. A much more recent application is being piloted by the U.S. Navy to use a much more complex simulation to immerse veterans suffering from PTSD in simulations of urban combat settings. Much as

in phobia treatment, exposure to the subject of the trauma or fear leads to desensitization, and a significant reduction in symptoms.

Other research fields in which the use of virtual reality is being explored are physical medicine, rehabilitation, physical therapy, and occupational therapy. In adult rehabilitation, a variety of virtual reality applications are currently being evaluated within upper and lower limb motor rehabilitation for individuals recovering from stroke or spinal cord injury. In pediatrics, the use of virtual reality is being evaluated to promote movement abilities, navigational abilities, or social skills in children with cerebral palsy, acquired brain injury, or other disabilities. Research evidence is emerging rapidly in the field of virtual reality for therapeutic uses. A number of recent reviews published in peer-reviewed journals have summarized the current evidence for the use of Virtual Reality within pediatric and adult rehabilitation. One such review concluded that the field is potentially promising.

Implementation

To develop a real time virtual environment, a computer graphics library can be used as embedded resource coupled with a common programming language, such as C++, Perl, Java, or Python. Some of the most popular computer graphic libraries are OpenGL, Direct3D, Java3D, and VRML, and their use are directly influenced by the system demands in terms of performance, program purpose, and hardware platform. The use of multithreading can also accelerate 3D performance and enable cluster computing with multi-user interactivity.

Manufacturing

Virtual reality can serve to new product design, helping as an ancillary tool for engineering in manufacturing processes, new product prototypes, and simulation. Among other examples, Electronic Design Automation, CAD, Finite Element Analysis, and Computer Aided Manufacturing are widely utilized programs. The use of Stereolithography and 3D printing shows how computer graphic modeling can be applied to create physical parts of real objects used in naval, aerospace, and automotive industries, which can be seen, for example, in the VR laboratory of VW in Mladá Boleslav. Beyond modeling assembly parts, 3D computer graphics techniques are currently used in the research and development of medical devices for therapies, treatments, patient monitoring, and early diagnoses of complex diseases.

Urban design

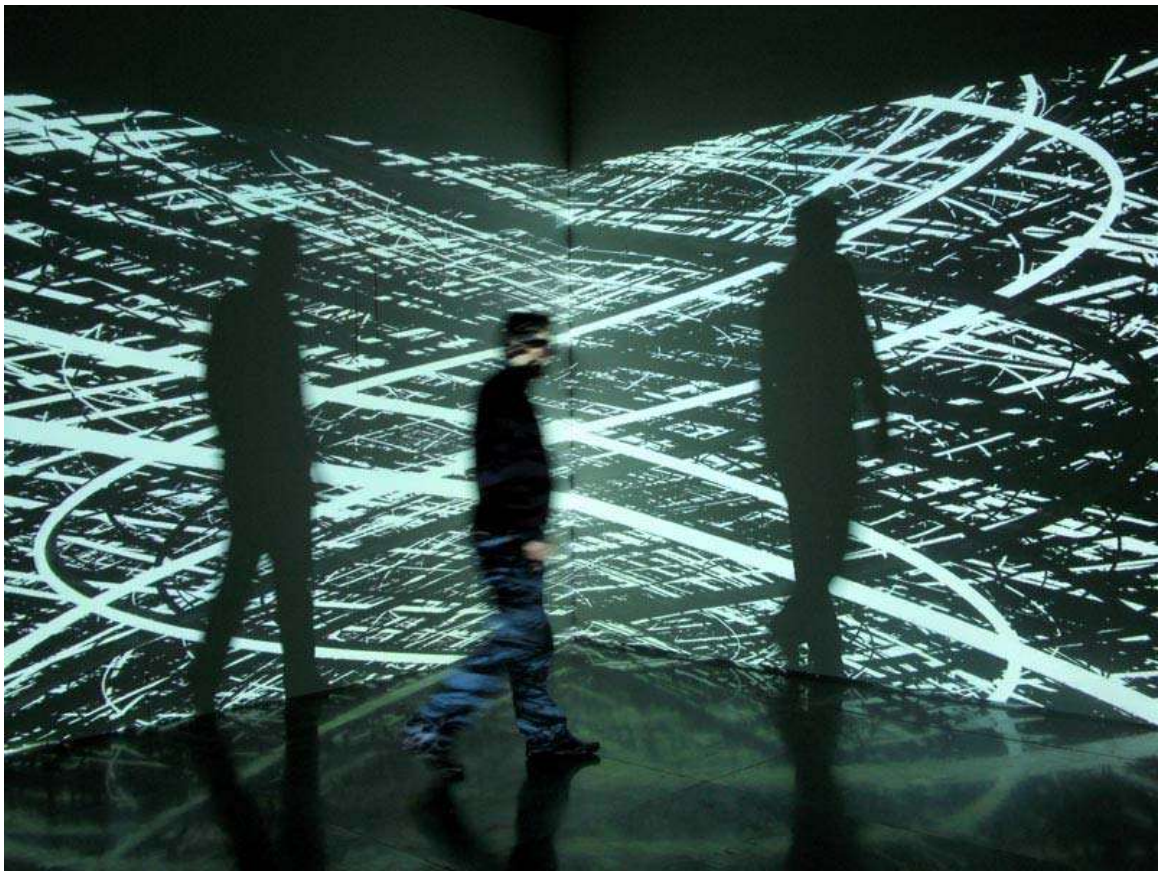
3D virtual reality is becoming widely used for urban regeneration and planning and transport projects.

Pioneers and notables

- Maurice Benayoun
- Mark Bolas
- Fred Brooks
- Edmond Couchot
- James H. Clark
- Doug Church
- Char Davies
- Tom DeFanti
- David Em
- Scott Fisher
- William Gibson
- Morton Heilig
- Myron Krueger
- Knowbotic Research
- Jaron Lanier
- Brenda Laurel
- Michael Naimark
- Randy Pausch
- Mark Pesce
- Susumu Tachi
- Ivan Sutherland

Chapter 15

Immersion



Pascal Dombis Installation view of *Irrationnal Geometrics* 2008

Immersion is the state of consciousness where an immersant's awareness of physical self is diminished or lost by being surrounded in an engrossing total environment; often artificial. This mental state is frequently accompanied with spatial excess, intense focus, a distorted sense of time, and effortless action. The term is widely used for describing immersive virtual reality, installation art and video games, but it is not clear if people are using the same word consistently. The term is also cited as a frequently-used buzzword, in which case its meaning is intentionally vague, but carries the connotation of being particularly engrossing.

The sensation of **total immersion** in virtual reality (VR) can be described as implied complete presence within an insinuated space of a virtual surrounding where everything within that sphere relates necessarily to the proposed "reality" of that world's cyberspace and where the immersant is seemingly altogether disconnected from exterior physical space.

Types of immersion

Classic Virtual reality HMD

According to Ernest Adams, author and consulter on game design, immersion can be separated into three main categories:

Tactical immersion
: Tactical immersion is experienced when performing tactile operations that involve skill. Players feel "in the zone" while perfecting actions that result in success.

Strategic immersion
: Strategic immersion is more cerebral, and is associated with mental challenge. Chess players experience strategic immersion when choosing a correct solution among a broad array of possibilities.

Narrative immersion
: Narrative immersion occurs when players become invested in a story, and is similar to what is experienced while reading a book or watching a movie.

Staffan Björk and Jussi Holopainen, in *Patterns In Game Design*, divide immersion into similar categories, but call them **sensory-motoric immersion**, **cognitive immersion** and **emotional immersion**, respectively. In addition to these, they add three new categories:

Spatial immersion
: Spatial immersion occurs when a player feels the simulated world is perceptually convincing. The player feels that he or she is really "there" and that a simulated world looks and feels "real".

Psychological immersion
: Psychological immersion occurs when a player confuses the game with real life.

Sensory immersion
: The experience of entering into the three-dimensional environment, and being intellectually stimulated by it. The player experiences a unity of time and space as the player fuses with the image medium, which affects impression and awareness.

Immersive virtual reality

The Cave Automatic Virtual Environment

Immersive virtual reality is a hypothetical future technology that exists today as virtual reality art projects, for the most part. It consists of immersion in an artificial environment where the user feels just as immersed as they usually feel in consensus reality.

Direct stimulation of the nervous system

The most considered method would be to induce the sensations that made up the virtual reality in the nervous system directly. In functionalism/conventional biology we interact with consensus reality through the nervous system. Thus we receive all input from all the senses as nerve impulses. It gives your neurons a feeling of heightened sensation. It would involve the user receiving inputs as artificially stimulated nerve impulses, the system would receive the CNS outputs (natural nerve impulses) and process them allowing the user to interact with the virtual reality. Natural impulses between the body and central nervous system would need to be prevented.

Requirements

Understanding of the nervous system

A comprehensive understanding of which nerve impulses correspond to which sensations, and which motor impulses correspond to which muscle contractions will be required. This will allow the correct sensations in the user, and actions in the virtual reality to occur. The Blue Brain Project is the current, most promising research with the idea of understanding how the brain works by building very large scale computer models.

Ability to manipulate CNS

The nervous system would obviously need to be manipulated. Whilst non-invasive devices using radiation have been postulated, invasive cybernetic implants are likely to become available sooner and be more accurate. Manipulation could occur at any stage of the nervous system - the spinal chord is likely to be simplest; as all nerves pass through here, this could be the only site of manipulation. Molecular Nanotechnology is likely to provide the degree of precision required and could allow the implant to be built inside the body rather than be inserted by an operation.

Computer hardware/software to process inputs/outputs

A very powerful and probably (but not necessarily) Strong AI would be required to process all the inputs from the CNS, run a simulation of a virtual reality approaching the complexity of consensus reality, and translate its events to a complete set of nerve impulses for the user. Strong artificial intelligence may also be required to write the program for a decent alternate reality.

Immersive digital environments

Cosmopolis (2005), Maurice Benayoun's Giant Virtual Reality Interactive Installation

An **immersive digital environment** is an artificial, interactive, computer-created scene or "world" within which a user can immerse themselves.

Immersive digital environments could be thought of as synonymous with Virtual reality, but without the implication that actual "reality" is being simulated. An immersive digital environment could be a model of reality, but it could also be a complete fantasy user interface or abstraction, as long as the user of the environment is immersed within it. The definition of immersion is wide and variable, but here it is assumed to mean simply that the user feels like they are part of the simulated "universe". The success with which an immersive digital environment can actually immerse the user is dependent on many factors such as believable 3D computer graphics, surround sound, interactive user-input and other factors such as simplicity, functionality and potential for enjoyment. New technologies are currently under development which claim to bring realistic environmental effects to the players' environment - effects like wind, seat vibration and ambient lighting.

Perception

To create a sense of full immersion, the 5 senses (sight, sound, touch, smell, taste) must perceive the digital environment to be physically real. Immersive technology can perceptually fool the senses through:

- Panoramic 3D displays (visual)
- Surround sound acoustics (auditory)
- Haptics and force feedback (tactile)
- Smell replication (olfactory)
- Taste replication (gustation)

Interaction

Once the senses reach a sufficient belief that the digital environment is real, the user must then be able to interact with the environment in a natural, intuitive manner. Various immersive technologies such as gestural controls, motion tracking, and computer vision respond to the user's actions and movements. Brain control interfaces (BCI) respond to the user's brainwave activity.

Examples and applications

Computer games from simple arcade to Massively multiplayer online game and training programs such as flight and driving simulators. Entertainment environments such as motion simulators that immerse the riders/players in a virtual digital environment enhanced by motion, visual and aural cues. There is a motion simulators of the Virunga Mountains in Rwanda to meet a tribe of Mountain Gorillas, or a ride that takes a journey through the arteries and heart to witness the build up of plaque and thus learn about cholesterol and health.

There are art installations such as those made by Knowbotic Research, Donna Cox, Rebecca Allen, Maurice Benayoun, Char Davies, StudioIMC and Jeffrey Shaw.

More generic examples of an IDE include:

- Any computer application or software program.
- Interactive TV shows or services such as CNN text.
- A VoIP conversation or chat session.
- A physical environment / immersive space with surrounding digital projections and sound such as the CAVE
- The use of head-mounted displays for viewing movies, with head-tracking and computer control of the image presented, so that the viewer appears to be inside the scene.

To some extent screensavers and DVD movies are also immersive digital environments, though usually not interactive.

Chapter 16

Futures Techniques

In the multi-disciplinary field of futurology, futurologists use a diverse range of forecasting methods, including:

Anticipatory thinking protocols

Delphi method

The Delphi method is a very popular technique used in Futures Studies. It was developed by Gordon and Helmer in 1953 at RAND. It can be defined as a method for structuring a group communication process, so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem.

It uses the iterative, independent questioning of a panel of experts to assess the timing, probability, significance and implications of factors, trends and events in the relation to the problem being considered. Panelists are not brought together but individually questioned in rounds. After the initial round, the panelists are given lists of anonymous answers from other panelists which they can use to refine their own views.

Studies employing Delphi method tend to be difficult to perform. The application of the Delphi method requires a great deal of attention to the selection of participating experts and the questionnaires have to be scrupulously prepared and tested in advance. The initial preparation and follow-up rounds of questioning of the panelists tends to be time consuming.

Delphi's primary strength is its ability to explore, tranquilly and objectively, issues that require judgement. Unlike panel sessions, the iterative Delphi method, allows the forecasting and assessment to be done without the effect of strong personalities or reputations influencing other panelists and also overcomes the difficulty of getting all experts together in a single time and place.

Causal layered analysis (CLA)

This method, developed by Sohail Inayatullah, is one of the newest developments in the Futures Studies. Causal layered analysis focuses on "opening up" the present and past to create alternative futures rather than on developing a picture of particular future.

It is concerned with the vertical dimension of futures studies, with the layers of analysis. CLA is based on the assumption that the way in which a problem is formulated changes the policy solutions and the actors in charge of initiating transformations.

The key principle of the method is using and integrating different ways of knowing.

There are a number of benefits arising from the application of this method. Causal layered analysis increases the range and richness of scenarios; leads to inclusion of different ways of knowing among participants in workshops; appeals to wider range of individuals through incorporation of non-textual and artistic elements; extends the discussion beyond the obvious to the deeper and marginal; and leads to the policy actions that can be educated by alternative layers of analysis.

Environmental scanning

Environmental scanning is usually used at the start of a futures project. It aims at broad exploration of all major trends, issues, advancements, events and ideas across a wide range of activities. Information is collected from many different sources, such as newspapers, magazines, Internet, television, conferences, reports, and also science-fiction books. Various tools and methodologies are used by large corporations to systematically scope their external environment. An example is the widely used FUTURE structure developed by Futurist Patrick Dixon described in his book Futurewise - F ast, U rban, T ribal, U niversal, R adical, E thical. Attention needs to be given to potential Wild Cards - low probability but potentially high impact events.

Scanning is used to build up a comprehensive picture of factors that could impact strategy.

Four types of indicators can be examined in the process of environmental scanning:

1. Lone signals (individual factors that might indicate change)
2. Landmark events (in various areas of life)
3. Forecasts of experts
4. Statistical descriptions (to portray development of elements of the study).

Morphological analysis

Morphological analysis is a technique developed by Fritz Zwicky (1966, 1969) for exploring all the possible solutions to a multi-dimensional, non-quantified problem complex.

As a problem structuring and problem solving technique, morphological analysis was designed for multi-dimensional, non-quantifiable problems where causal modeling and simulation do not function well or at all. Zwicky developed this approach to address seemingly non-reducible complexity. Using the technique of cross consistency assessment (CCA), the system however does allow for reduction, not by reducing the

number of variables involved, but by reducing the number of possible solutions through the elimination of the illogical solution combinations in a grid box.

MA has also been employed for the identification of new product opportunities. The technique involves mapping options in order to attain an overall perspective of possible solutions. It comprises the two main activities: a systematic analysis of a current and future structure of the area including the gaps in that structure, stimulation for creation of new alternative, which could fill the gaps and meet any needs.

Scenario planning

Scenarios are one of the most popular and persuasive methods used in the Futures Studies. Government planners, corporate strategists and military analysts use them in order to aid decision-making. The term scenario was introduced into planning and decision-making by Herman Kahn in connection with military and strategic studies done by RAND in the 1950s.

It can be defined as a rich and detailed portrait of a plausible future world, one sufficiently vivid that a planner can clearly see and comprehend the problems, challenges and opportunities that such an environment would present.

A scenario is not a specific forecast of the future, but a plausible description of what might happen. Scenarios are like stories built around carefully constructed plots based on trends and events. They assist in selection of strategies, identification of possible futures, making people aware of uncertainties and opening up their imagination and initiating learning processes.

One of the key strengths of the scenario process is its influence on the way of thinking of its participants. A mindset, in which the focus is placed on one possible future, is altered towards the balanced thinking about a number of possible alternative futures.

Although it is a very rewarding method it is also very demanding. The difficulties in its use can arise from a lack of clear focus, purpose or directions. As a result too many scenario stories can be created and/or their content may not be directly related to the strategic question.

Future history

A **future history** is a postulated history of the future.Some science fiction authors construct as a common background for fiction. The author may include a timeline of events for this history. A related field is alternate history, which assumes that the events at a critical point in past history turned out differently and then draws a fictional future timeline from that event.

Monitoring

It is a process that aims at evaluation of events, as they occur or just after. It involves activities like scanning, detecting, projecting, assessing, responding and tracking. Monitoring is one of the fundamental activities performed by Futures Studies.

Content analysis

This technique is used for the systematic and objective study of the particular aspects of various 'messages'. Such 'messages' can be found in books, journals, newspapers, private letters, publications of political parties, reports, surveys, interviews, television, Internet and so on. This method, in order to be reliable and valid, needs to be performed with high competency.

Backcasting (eco-history)

It is a technique that often is pointed out as an opposite to forecasting. It involves identification of a particular scenario and tracing its origins and lines of development back to the present.

Back-view mirror analysis

It builds upon the assumption that any future oriented group process has to manage peoples' difficulties in thinking into the future. These difficulties can arise from the fears as well as from the lack of experience in futures thinking.

Back-view mirror analysis allows dealing with the fears related to the future by creating a new perspective that looks to the past instead of starting the process in the present. The method is used to perform qualitative analysis of the past using both quantitative and qualitative data.

Cross-impact analysis

The method was developed by Theodore Gordon and Olaf Helmer in 1966 in an attempt to answer a question whether perceptions of how future events may interact with each other can be used in forecasting.

As it is well known, most events and trends are interdependent in some ways. Cross-impact analysis provides an analytical approach to the probabilities of an element in a forecast set, and it helps to assess probabilities in view of judgements about potential interactions between those elements.

The technique can be used by individuals and groups at an elementary qualitative level as well as it can be employed to perform more complicated and intensive quantitative analysis. One of its strengths is that it forces the attention towards "chains of causality: x affects y; y affects z". On the other side it can be very fatiguing and monotonous.

Futures workshops

Future workshops were developed by Robert Jungk in order to allow anybody to become involved in creating their preferred future rather than being subjected to decisions made by experts. Future workshops are very strongly action oriented. They aim, first to imagine the desired future, and then to plan it and implement it.

Future workshops have four distinctive phases:

1. In the first preparatory phase the issue that will be considered is identified and the structure and details of sessions are arranged.
2. The operative phase involves clarification of the issue considered and articulation of negative experiences in the present situation.
3. In the fantasy stage participants verbalise their desires, dreams, fantasies and views about the future in a free idea generation session. The participants are asked to forget all the limitations and obstacles of the present reality.
4. The last step involves: analysis of the feasibility of ideas and solutions generated in the fantasy phase; recognition of limits and barriers for implementation and discovering how they can be overcome.

Failure mode and effects analysis

Failure mode and effects analysis (FMEA) is a safety analysis method first developed for systems engineering which examines potential failures in products or processes. It may be used to evaluate risk management priorities for mitigating known threat-vulnerabilities.

FMEA helps select remedial actions that reduce **cumulative impacts** of life cycle consequences (risks) from a systems failure (fault).

Measured Action

Measured Action is an action that improves Production Capability by a measured amount, (e.g. if X hours of overtime then Y additional files transmitted). Measured actions enhance FMEA by perfectly matching calculated risks with calculated contingencies to ensure a specific outcome.

Futures biographies

This method, also called futures imagining, aims to create individual imaginaries, to gather peoples' views on the future and to examine them in the study of collective future. Peoples' expectations and opinions are considered as an important indication of possible goals and to possible directions that can influence peoples' actions and in result steer the future.

Futures wheel

The method is a form of structured brainstorming that aims at identifying and packaging secondary and tertiary consequences of trends and events. A trend or event is placed in the middle of a piece of paper and then small spokes are drawn wheel-like from the centre. Primary impacts and consequences are written in circles of the first ring.

Then secondary consequences of each primary impact are derived forming the second ring. This ripple effect continues until there is a clear picture of implications that the event or trend can have. Futures wheel is a very simple but powerful technique for drawing out people's opinions and ideas. However, it is sensitive to underlying assumptions.

Relevance tree

It is an analytical technique that subdivides a large subject into increasingly smaller subtopics. The relevance tree has a form of a hierarchical structure that begins with a high level of abstraction and moves down with greater degree of detail in the following levels of the tree. It is a powerful technique that helps to ensure that a given problem or issue is broken into comprehensive detail and that important connections among the elements considered are presented in both current and potential situations.

Simulation and modelling

Simulation and modelling are computer-based tools developed to represent reality. They are widely used to analyse behaviours and to understand processes. Models allow demonstration of past changes as well as the examination of various transformations and their impact on each other and other considered factors.

They can help to understand the connections between factors and events and to examine their dynamics. Simulation is a process that represents a structure and change of a system. In simulation some aspects of reality are duplicated or reproduced, usually within the model. The main purpose of simulation is to discern what would really happen in the real world if certain conditions, imitated by the model, developed.

Although modelling and simulation became even more popular with the development of computing technology, application of these techniques have certain limits. Models represent a simplification of a system that is being examined; therefore the results need to be carefully considered.

As the complexity of real systems increases models need to be more and more complex to represent the reality most accurately. In result, they may become increasingly difficult to understand and to be operated. Their complex nature can cause problems with using and managing results.

As models constitute a simpler version of reality, certain factors can be omitted, and in consequence can lead to mistakes. Such mistakes are not easy to be found and corrected.

Social network analysis

Social network analysis (also sometimes called *network theory*) has emerged as a key technique in modern sociology, anthropology, Social Psychology and organizational studies, as well as a popular topic of speculation and study.

Research in a number of academic fields have demonstrated that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals.

Systems engineering

An interdisciplinary approach to engineering systems that is inherently complex, since the behavior of and interaction among system components is not always well defined.

Defining and characterizing such systems and subsystems, and the interactions among them, is the primary aim of systems engineering. On *very* large programs, a systems architect may be designated to serve as an interface between the user/sponsor and systems engineer.

RADM Grace Hopper, USNR was quoted as saying "Life was simple before World War II. After that, we had systems."

Visioning

Visioning is a popular method in the studies of desirable futures and the one that gives emphasis to values. It is extensively used in urban planning. The visioning process is based on the assumption that images of the future lead peoples' present behaviours, guide choices and influence decisions. Images of the future can be positive or negative and cause different responses according to the perceptions.

Vision is usually seen as a positive, desirable image of the future and can be defined as a compelling, inspiring statement of the preferred future that the authors and those who subscribe to the vision want to create.

There are a number of issues that need to be addressed while using the visioning method. Vision comprises peoples' values, wishes, fears and desires. In order to make the visioning process work it is necessary to ensure that it is not making an idealistic wish-list; that vision is an image of the future shared by a whole community; and that the vision is translatable into reality.

Trend analysis

Trend analysis is one of the most often used methods in forecasting. It aims to observe and register the past performance of a certain factor and project it into the future. It involves analysis of two groups of trends: quantitative, mainly based on statistical data, and qualitative, these are at large concerned with social, institutional, organisational and political patterns.

In the quantitative trend analysis data is plotted along a time axis, so that a simple curve can be established. Short term forecasting seems quite simple; it becomes more complex when the trend is extrapolated further into the future, as the number of dynamic forces that can change direction of the trend increases. This form of simple trend extrapolation helps to direct attention towards the forces, which can change the projected pattern.

A more elaborated curve that uses times series analysis can often reveal surprising historical and current data patterns. The qualitative trend analysis is one of the most demanding and creative methods in Futures Studies.

As trends never speak for themselves, the identification and description of patterns is partly empirical and partly creative activity. The most challenging part of qualitative trends analysis is identification of a tendency early, as recognition of a mature trend is "relatively useless" in influencing anyone's behaviour.

Adaptive role-playing

Although similar to decision theory, game theory studies decisions that are made in an environment where various players interact. In other words, game theory studies choice of optimal behavior when costs and benefits of each option are not fixed, but depend upon the choices of other individuals.

ALL-WinWin collaborative efforts require group decision support systems (GDSS) that enable the knowledge management community of practice to assure sustainable mutually-beneficial results.

Chapter 17

Technology Roadmap

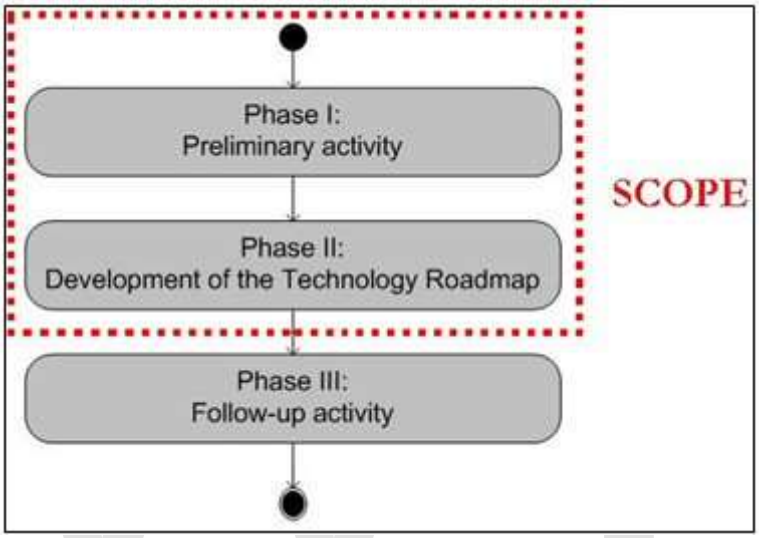


Fig 1: The Technology Roadmapping phases

A **technology roadmap** is a plan that matches short-term and long-term goals with specific technology solutions to help meet those goals. It is a plan that applies to a new product or process, or to an emerging technology. Developing a roadmap has three major uses. It helps reach a consensus about a set of needs and the technologies required to satisfy those needs; it provides a mechanism to help forecast technology developments and it provides a framework to help plan and coordinate technology developments.

The existence of product managers in the product software industry indicates that software is becoming more commercialized as a standard product. This manager is responsible over the whole line of software requirement management, defining of products and their releases and this with all internal and external stakeholders involved. In this context, product roadmapping can be placed to aid software product managers in planning and placing their products with the use of scientific and technological resources. For managing and using the technological resources technology planning can be used.

The Roadmapping process

The Technology Roadmapping Process conducts 3 phases (see figure 1.): preliminary activities, the development of the roadmap and the follow-up activities phase. Because the process is too big for one model the phases are modeled separately. Only the first two phases are considered. In the models no different roles are made, this is because everything is done by the participants as a group.

Phase 1: Preliminary phase

The first phase, the preliminary phase (see figure 2.), consists of 3 steps: *satisfy essential conditions, provide leadership / sponsorship and define the scope and boundaries for the technology roadmap*. In this phase the key decision makers must identify that they have a problem and that technology roadmapping can help them in solving the problem.

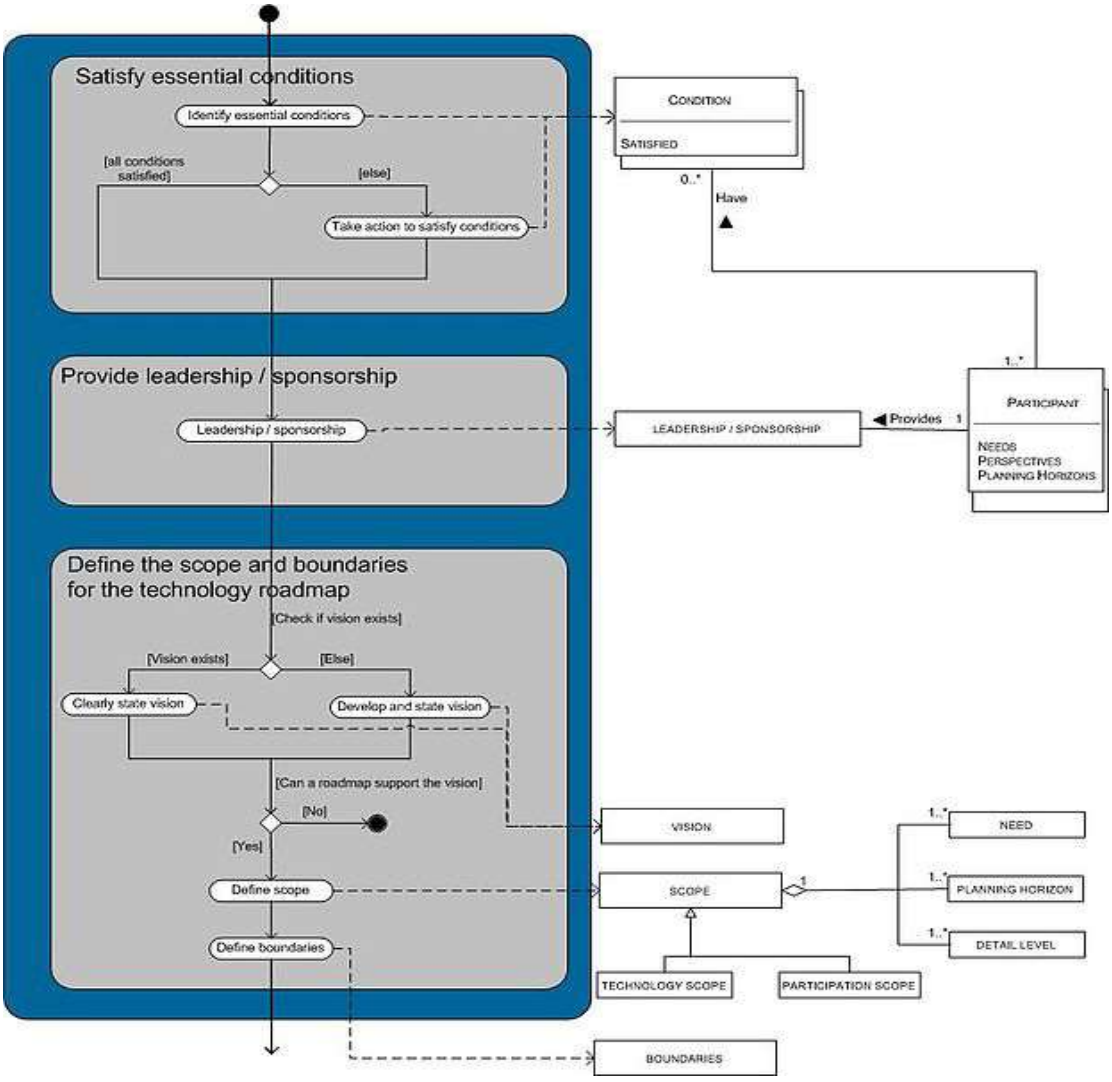


Figure 2. The process-data model of the preliminary phase

Satisfy essential conditions

In this step it must become clear what the conditions are (they have to be identified) and if they are not met that somebody will take the actions necessary to meet the unmet conditions. These conditions include for example the following: there must be a need for the technology roadmap, input and participation from several different parts of the organization (e.g. marketing, R&D, the Strategic Business Units) with different planning horizons and different perspectives and the process should be needs driven. All the conditions should be satisfied (or someone is going to take the actions necessary) in order to continue to the next step. The participants can have zero or more conditions of their own. It applies to all the conditions that they have the attribute to be met or not.

Provide leadership / sponsorship

Committed leadership is needed because time and effort is involved in creating the technology roadmap. Additionally the leadership should come from one of the participants, one of them provides leadership / sponsorship. This means that the line organization must drive the process and use the roadmap to make resource allocation decisions.

Define the scope and boundaries for the technology roadmap

In this step the context for the roadmap will be specified. In the company a vision should exist and it must be clear that the roadmap can support that vision. If the vision does not exist one should be developed and clearly stated. When that is done the boundaries and the scope of the roadmap should be specified. Furthermore the planning horizon and the level of details should be set. The scope can be further divided into the technology scope and the participation scope.

In table 1. all the different sub-activities of the preliminary activity phase can be seen. All the sub-activities have concepts as end "products", these are marked in bold. These concepts are the actual meta-data model, which is an adjusted class diagram.

Table 1. Activity table for the preliminary activity phase

Activity	Sub-Activity	Description
Satisfy essential conditions	Identify essential conditions	When all the **participants** come together, essential **conditions**, like what groups should be involved, what are the key customers and what are the key suppliers, can be identified.
	Take action to satisfy conditions	For technology roadmapping to succeed, **conditions** from the **participants** must be satisfied.
Provide leadership / sponsorship		The part of **leadership** / **sponsorship** should be taken by line organization; they must drive the roadmapping process and use the roadmap to

		make resource allocation decisions.
Define the scope and boundaries for the technology roadmap	Clearly state vision	The already existing **vision** has to be clear.
	Develop vision	The **vision** is developed and stated clearly.
	Define scope	The **scope** of the project can further define the set of **needs**, **planning horizon** and level of **detail**. The scope can be further divided into the **technology scope** and the **participation scope**.
	Define boundaries	The **boundaries** should also be included.

Phase 2: Development phase

The second phase, the development of the technology roadmap phase (see figure 3.), consists of 7 steps: *identify the "product" that will be the focus of the roadmap, identify the critical system requirements and their targets, specify the major technology areas, specify the technology drivers and their targets, identify technology alternatives and their timelines, recommend the technology alternatives that should be pursued and create the technology roadmap report*. These steps create the actual roadmap.

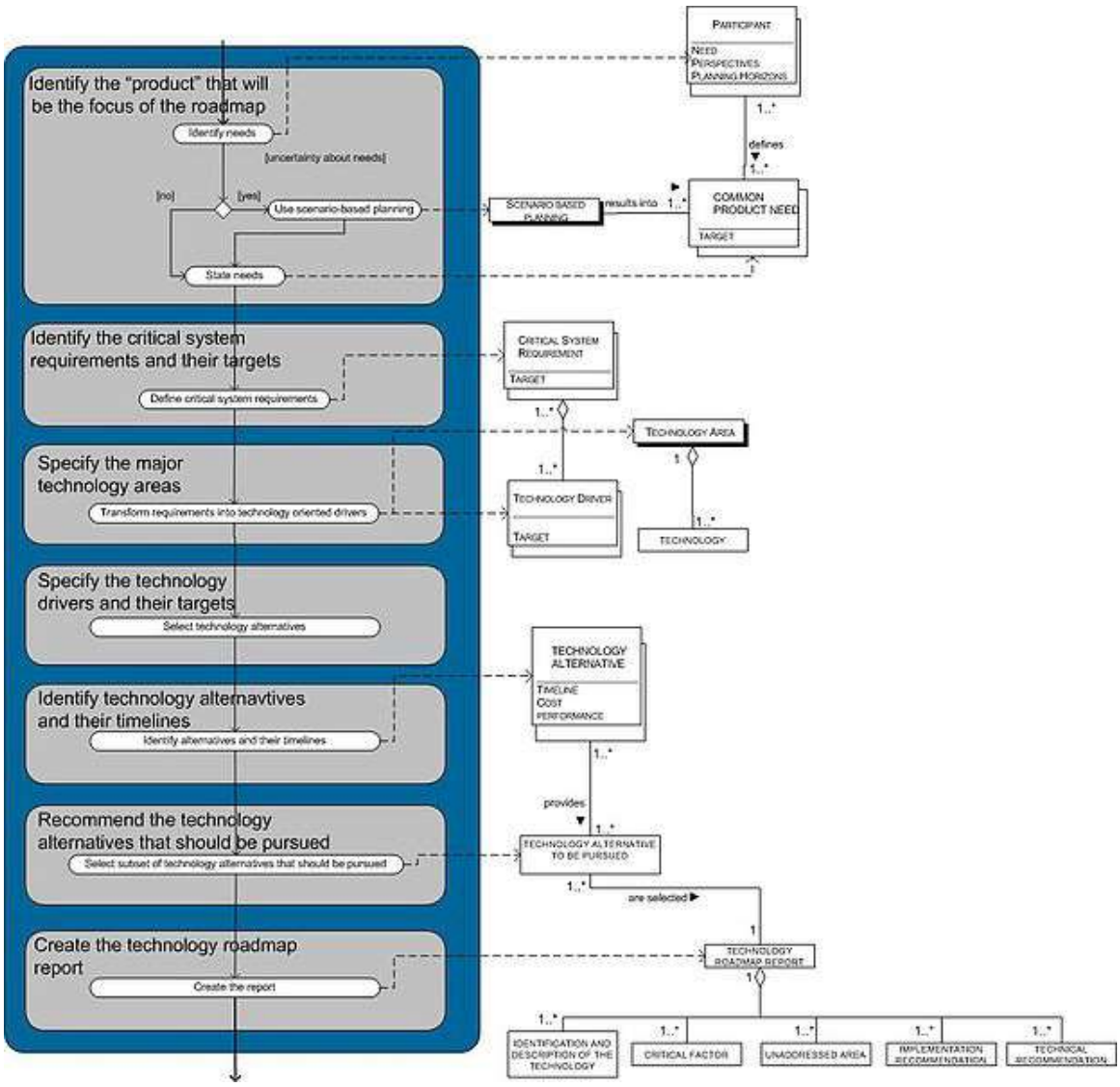


Figure 3. The process-data model of the development phase

Identify the "product" that will be the focus of the roadmap

In this step the common product needs are identified and should be agreed on by all the participants. This is important to get the acceptance of all groups for the process. In case of uncertainty of the product needs scenario-based planning can be used to determine the common product needs. In figure 3. it can be seen that the participants and possibly the scenario-based planning provide the common product needs.

Identify the critical system requirements and their targets

Once it is decided what needs to be roadmapped the critical system requirements can be identified, they provide the overall framework for the technology roadmap. The requirements can have targets (as an attribute in figure 3.) like reliability and costs.

Specify the major technology areas

These are the areas which can help achieve the critical system requirements. For each technology area several technologies can be found. Example technology areas are: Market assessment, Crosscutting technology, Component development and System development.

Specify the technology drivers and their targets

In this step the critical system requirements from step *Identify the critical system requirements and their targets* are transformed into technology drivers (with targets) for the specific technology area. These drivers are the critical variables that will determine which technology alternatives are selected. The drivers depend on the technology areas but they relate to how the technology addresses the critical system requirements.

Identify Technology alternatives and their timelines

At this point the technology drivers and their targets are specified and the technology alternatives that can satisfy those targets should be specified. For each of the alternatives a timeline should be estimated for how it will mature with respect to the technology driver targets.

Time

This factor can be adapted suitable for the particular situation. The time horizons for E-commerce and software related sectors are usually short. Other distinctions can be made on scale and intervals.

Recommend the technology alternatives that should be pursued

Because the alternatives may differ in costs, timeline etc. a selection has to be made of the alternatives. These will be the alternatives to be pursued in figure 3. In this step a lot of trade-off has to be made between different alternatives for different targets, performance over costs and even target over target.

Create the technology roadmap report

At this point the technology roadmap is finished. In figure 3, it can be seen that the technology roadmap report consists of 5 parts: the identification and description of each technology area, critical factors in the roadmap, unaddressed areas, implementation recommendations and technical recommendations. The report can also include additional information. In table 2. all the different sub-activities of the development phase can be seen.

Table 2. Activity table for the Development phase.

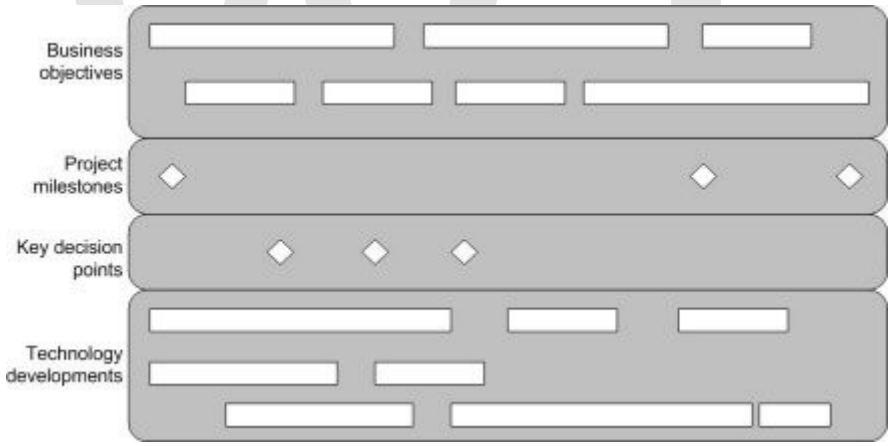
Activity	Sub-Activity	Description
Identify the "product" that will be the focus of the roadmap	Identify needs	This critical step is to get the participants to identify and agree on the COMMON PRODUCT NEEDS. This is important to get their buy-in and acceptance.
	Use Scenario-based planning	If there is major uncertainty about the COMMON PRODUCT NEEDS SCENARIO-BASED PLANNING can be used. Each scenario must be reasonable, internally consistent and comparable with the other scenarios.
	State needs	These are the NEEDS for the product.
Identify the critical system requirements and their targets	Define critical system requirements	The CRITICAL SYSTEM REQUIREMENTS provide the overall framework for the roadmap and are high-level dimensions to which the technologies relate. These include things like reliability and costs.
	Define targets	For each of the system requirements TARGETS have to be defined.
Specify the major technology areas	Transform requirements into technology oriented drivers	The major TECHNOLOGY AREAS should be specified to help achieve the CRITICAL SYSTEM REQUIREMENTS for the product. The CRITICAL SYSTEM REQUIREMENTS are then transformed into TECHNOLOGY DRIVERS for the specific TECHNOLOGY AREAS.
Specify the technology drivers and their targets	Select technology alternatives with their targets	TECHNOLOGY DRIVERS and their TARGETS are set based on the CRITICAL SYSTEM REQUIREMENT TARGETS. It specifies how viable TECHNOLOGY ALTERNATIVES must be to perform by a certain date. From the available TECHNOLOGY ALTERNATIVES a selection has to be made.
Identify technology alternatives and their timelines	Identify alternatives and their timelines	The TECHNOLOGY ALTERNATIVES that can satisfy the TARGETS must be identified. Next to this the TIMELINE from each alternative has to be identified.
Recommend the technology alternatives that should be	Select subset of technology alternatives to be pursued	Determine which TECHNOLOGY ALTERNATIVE TO PURSUE and when to shift to a different TECHNOLOGY. Consolidate the best information and

pursued		develop consensus from many experts.
Create the technology roadmap report	Create the report	Here the actual TECHNOLOGY ROADMAP REPORT is created. This report includes: IDENTIFICATION AND DESCRIPTION THE TECHNOLOGY, CRITICAL FACTOR, UNADDRESSED AREA, and IMPLEMENTATION RECOMMENDATION AND TECHNICAL RECOMMENDATION.

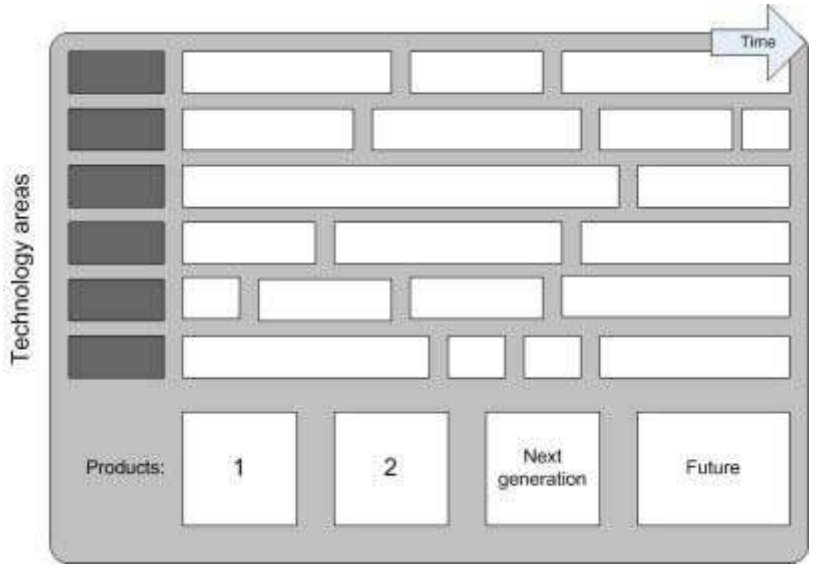
Phase 3: Follow-up activity phase

This is the moment when the roadmap must be critiqued, validated and hopefully accepted by the group that will be involved in any implementation. For this a plan needs to be developed using the technology roadmap. Next there must be a periodical review and update point, because the needs from the participants and the technologies are evolving.

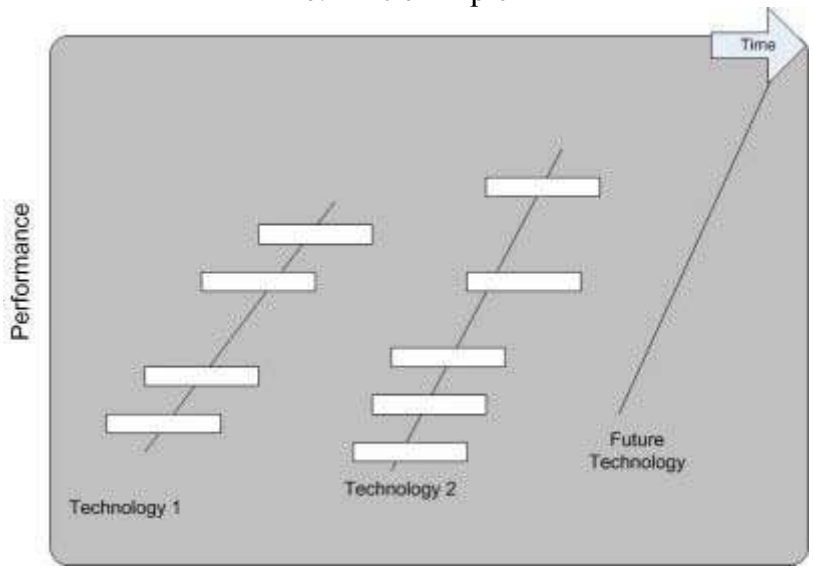
Planning and Business Development Context for Technology Roadmapping



5. Programme planning example



6. Bars example



7. Graphs example

The process of technology roadmapping fits into corporate strategy, corporate strategic planning, technology planning and the business development context. 3 critical elements should be connected: needs, products and technology.

Knowledge and skills required

In order to create a technology roadmap it is required to have a certain set of knowledge and skills. This means that some of the participants must know the process of technology roadmapping. Next to this group-process and interpersonal skills are required since the process includes a lot of discussions and finding out what the common need is. If the amount of participants is really large there might be need for a consultant or facilitator.

The purpose of technology Roadmapping

Product planning

This is the most common type of a technology roadmap; linking the insertion of technologies into products.

Programme planning

This type is more directed to the implementation of strategy and related to project planning. Figure 5 shows the relationships between technology development phases, programme phases and milestones.

The formats of technology Roadmapping

- Bars: Almost all the roadmaps are (partly) expressed in bars for each layer. This makes the roadmaps very simple and unified, which makes the communication and integration easier.
- Graphs: Also a technology roadmap can be expressed as a graph, usually one for each of the sub layers. (e.g. IMEC uses the second method).

Chapter 18

5G Technology

5G (*5th generation mobile networks* or *5th generation wireless systems*) is a name used in some research papers and projects to denote the next major phase of mobile telecommunications standards beyond the upcoming 4G standards (expected to be finalized between approximately 2011 and 2013). Currently, 5G is not a term officially used for any particular specification or in any official document yet made public by telecommunication companies or standardization bodies such as 3GPP, WiMAX Forum or ITU-R. New standard releases beyond 4G are in progress by standardization bodies, but at this time are not considered as new mobile generations since implementation and rollout of systems compliant with 4G is still under way; the goals of a 5G-based telecommunications network would ideally answer the challenges that a 4G model would present once it has entered widespread use.

Prognoses

The implementation of standards under a 5G umbrella would likely be around the year of 2020. A new mobile generation has appeared every 10th year since the first 1G system (NMT) was introduced in 1981, including the 2G (GSM) system that started to roll out in 1992, and 3G (W-CDMA/FOMA), which appeared in 2001. The development of the 2G (GSM) and 3G (IMT-2000 and UMTS) standards took about 10 years from the official start of the R&D projects, and development of 4G systems started in 2001 or 2002.

It is expected that in terms of data streams, a 5G standard would have peak download and upload speeds of more than the 1 Gbps to be offered by ITU-R's definition of 4G systems. The development of the bit rates offered by cellular systems is however hard to predict, since the historical bit rate development has shown very little resemblance with a simple exponential function of time (as opposed to for example Moore's law for computing capacity). The data rate increased by a factor 8 from 1G (NMT 1.2 kbps) to 2G (GSM 9.6 kbps). The peak bit rate increased by a factor 40 from 2G to 3G for mobile users (384 kbps), and by a factor of 200 from 2G to 3G for stationary users (2 Mbps). The peak bit rates are expected to increase by a factor 260 from 3G to 4G for mobile users (100 Mbps) and by a factor 500 from 3G to 4G for stationary users (1 Gbps).

Research

Key concepts suggested in research papers discussing 5G and beyond 4G wireless communications are:

- Pervasive networks providing *ubiquitous computing*: The user can simultaneously be connected to several wireless access technologies and seamlessly move between them. These access technologies can be 2.5G, 3G, 4G, or 5G mobile networks, Wi-Fi, WPAN, or any other future access technology. In 5G, the concept may be further developed into multiple concurrent data transfer paths.
- Cognitive radio technology, also known as smart-radio: allowing different radio technologies to share the same spectrum efficiently by adaptively finding unused spectrum and adapting the transmission scheme to the requirements of the technologies currently sharing the spectrum. This dynamic radio resource management is achieved in a distributed fashion, and relies on software defined radio.
- Internet protocol version 6 (IPv6), where a visiting care-of mobile IP address is assigned according to location and connected network.
- High altitude stratospheric platform station (HAPS) systems.
- *Real wireless world* with no more limitation with access and zone issues.
- Wearable devices with AI capabilities.
- One unified global standard.

The radio interface of 5G communication systems is suggested in a Korean research and development program to be based on beam division multiple access (BDMA) and group cooperative relay techniques.

Ubiquitous computing

Ubiquitous computing (**ubicomp**) is a post-desktop model of human-computer interaction in which information processing has been thoroughly integrated into everyday objects and activities. In the course of ordinary activities, someone "using" ubiquitous computing engages many computational devices and systems simultaneously, and may not necessarily even be aware that they are doing so. This model is usually considered an advancement from the desktop paradigm. More formally Ubiquitous computing is defined as "machines that fit the human environment instead of forcing humans to enter theirs."

This paradigm is also described as **pervasive computing**, ambient intelligence., where each term emphasizes slightly different aspects. When primarily concerning the objects involved, it is also **physical computing**, the *Internet of Things*, *haptic computing*, and *things that think*. Rather than propose a single definition for ubiquitous computing and for these related terms, a taxonomy of properties for ubiquitous computing has been

proposed, from which different kinds or flavors of ubiquitous systems and applications can be described.

Core concept

At their core, all models of ubiquitous computing (also called pervasive computing) share a vision of small, inexpensive, robust networked processing devices, distributed at all scales throughout everyday life and generally turned to distinctly common-place ends. For example, a domestic ubiquitous computing environment might interconnect lighting and environmental controls with personal biometric monitors woven into clothing so that illumination and heating conditions in a room might be modulated, continuously and imperceptibly. Another common scenario posits refrigerators "aware" of their suitably-tagged contents, able to both plan a variety of menus from the food actually on hand, and warn users of stale or spoiled food.

Ubiquitous computing presents challenges across computer science: in systems design and engineering, in systems modelling, and in user interface design. Contemporary human-computer interaction models, whether command-line, menu-driven, or GUI-based, are inappropriate and inadequate to the ubiquitous case. This suggests that the "natural" interaction paradigm appropriate to a fully robust ubiquitous computing has yet to emerge - although there is also recognition in the field that in many ways we are already living in an ubicomp world. Contemporary devices that lend some support to this latter idea include mobile phones, digital audio players, radio-frequency identification tags, GPS, and interactive whiteboards.

- Mark Weiser proposed three basic forms for ubiquitous system devices

These three forms proposed by Weiser are characterized by being macro-sized, having a planar form and on incorporating visual output displays. If we relax each of these three characteristics we can expand this range into a much more diverse and potentially more useful range of Ubiquitous Computing devices. Hence, three additional forms for ubiquitous systems have been proposed:

- *Dust*: miniaturized devices can be without visual output displays, e.g., Micro Electro-Mechanical Systems (MEMS), ranging from nanometres through micrometers to millimetres.
- *Skin*: fabrics based upon light emitting and conductive polymers, organic computer devices, can be formed into more flexible non-planar display surfaces and products such as clothes and curtains. MEMS device can also be painted onto various surfaces so that a variety of physical world structures can act as networked surfaces of MEMS.
- *Clay*: ensembles of MEMS can be formed into arbitrary three dimensional shapes as artefacts resembling many different kinds of physical object.

In his book *The Rise of the Network Society*, Manuel Castells suggests that there is an ongoing shift from already-decentralised, stand-alone microcomputers and mainframes

towards entirely pervasive computing. In his model of a pervasive computing system, Castells uses the example of the Internet as the start of a pervasive computing system. The logical progression from that paradigm is a system where that networking logic becomes applicable in every realm of daily activity, in every location and every context. Castells envisages a system where billions of miniature, ubiquitous inter-communication devices will be spread worldwide, "like pigment in the wall paint".

History

Mark Weiser coined the phrase "ubiquitous computing" around 1988, during his tenure as Chief Technologist of the Xerox Palo Alto Research Center (PARC). Both alone and with PARC Director and Chief Scientist John Seely Brown, Weiser wrote some of the earliest papers on the subject, largely defining it and sketching out its major concerns.

Recognizing that the extension of processing power into everyday scenarios would necessitate understandings of social, cultural and psychological phenomena beyond its proper ambit, Weiser was influenced by many fields outside computer science, including "philosophy, phenomenology, anthropology, psychology, post-Modernism, sociology of science and feminist criticism." He was explicit about "the humanistic origins of the ‘invisible ideal in post-modernist thought'", referencing as well the ironically dystopian Philip K. Dick novel Ubik.

MIT has also contributed significant research in this field, notably Hiroshi Ishii's *Things That Think* consortium at the Media Lab and the CSAIL effort known as Project Oxygen. Other major contributors include Georgia Tech's College of Computing, NYU's Interactive Telecommunications Program, UC Irvine's Department of Informatics, Microsoft Research, Intel Research and Equator, Ajou University UCRi & CUS.

Examples

One of the earliest ubiquitous systems was artist Natalie Jeremijenko's "Live Wire", also known as "Dangling String", installed at Xerox PARC during Mark Weiser's time there. This was a piece of string attached to a stepper motor and controlled by a LAN connection; network activity caused the string to twitch, yielding a *peripherally noticeable* indication of traffic. Weiser called this an example of *calm technology*.

Ambient Devices has produced an "orb", a "dashboard", and a "weather beacon": these decorative devices receive data from a wireless network and report current events, such as stock prices and the weather, like the Nabaztag produced by Violet Snowden.

Current research

Ubiquitous computing touches on a wide range of research topics, including distributed computing, mobile computing, sensor networks, human-computer interaction, and artificial intelligence.

Cognitive radio

Cognitive radio is a paradigm for wireless communication in which either a network or a wireless node changes its transmission or reception parameters to communicate efficiently avoiding interference with licensed or unlicensed users. This alteration of parameters is based on the active monitoring of several factors in the external and internal radio environment, such as radio frequency spectrum, user behaviour and network state.

History

The idea of cognitive radio was first presented officially by Joseph Mitola III in a seminar at KTH, The Royal Institute of Technology, in 1998, published later in an article by Mitola and Gerald Q. Maguire, Jr in 1999. It was a novel approach in wireless communications that Mitola later described as:

The point in which wireless personal digital assistants (PDAs) and the related networks are sufficiently computationally intelligent about radio resources and related computer-to-computer communications to detect user communications needs as a function of use context, and to provide radio resources and wireless services most appropriate to those needs.

It was thought of as an ideal goal towards which a software-defined radio platform should evolve: a fully reconfigurable wireless black-box that automatically changes its communication variables in response to network and user demands.

Regulatory bodies in various countries (including the Federal Communications Commission in the United States, and Ofcom in the United Kingdom) found that most of the radio frequency spectrum was inefficiently utilized. For example, cellular network bands are overloaded in most parts of the world, but amateur radio and paging frequencies are not. Independent studies performed in some countries confirmed that observation, and concluded that spectrum utilization depends strongly on time and place. Moreover, fixed spectrum allocation prevents rarely used frequencies (those assigned to specific services) from being used by unlicensed users, even when their transmissions would not interfere at all with the assigned service. This was the reason for allowing unlicensed users to utilize licensed bands whenever it would not cause any interference (by avoiding them whenever legitimate user presence is sensed). This paradigm for wireless communication is known as cognitive radio.

The first phone call over a cognitive radio network was made on Monday 11 January 2010 in Centre for Wireless Communications at University of Oulu using CWC's cognitive radio network CRAMNET (Cognitive Radio Assisted Mobile Ad Hoc Network), that has been developed solely by CWC researchers.

Terminology

Depending on the set of parameters taken into account in deciding on transmission and reception changes, and for historical reasons, we can distinguish certain types of cognitive radio. The main two are:

- **Full Cognitive Radio** ("Mitola radio"): in which every possible parameter observable by a wireless node or network is taken into account.
- **Spectrum Sensing Cognitive Radio**: in which only the radio frequency spectrum is considered.

Also, depending on the parts of the spectrum available for cognitive radio, we can distinguish:

- **Licensed Band Cognitive Radio**: in which cognitive radio is capable of using bands assigned to licensed users, apart from unlicensed bands, such as U-NII band or ISM band. The IEEE 802.22 working group is developing a standard for wireless regional area network (WRAN) which will operate in unused television channels.
- **Unlicensed Band Cognitive Radio**: which can only utilize unlicensed parts of radio frequency spectrum. One such system is described in the IEEE 802.15 Task group 2 specification. which focuses on the coexistence of IEEE 802.11 and Bluetooth.

Technology

Although cognitive radio was initially thought of as a software-defined radio extension (Full Cognitive Radio), most of the research work is currently focusing on Spectrum Sensing Cognitive Radio, particularly in the TV bands. The essential problem of Spectrum Sensing Cognitive Radio is in designing high quality spectrum sensing devices and algorithms for exchanging spectrum sensing data between nodes. It has been shown that a simple energy detector cannot guarantee the accurate detection of signal presence, calling for more sophisticated spectrum sensing techniques and requiring information about spectrum sensing to be exchanged between nodes regularly. Increasing the number of cooperating sensing nodes decreases the probability of false detection.

Filling free radio frequency bands adaptively using OFDMA is a possible approach. Timo A. Weiss and Friedrich K. Jondral of the University of Karlsruhe proposed a spectrum pooling system in which free bands sensed by nodes were immediately filled by OFDMA subbands.

Applications of Spectrum Sensing Cognitive Radio include emergency networks and WLAN higher throughput and transmission distance extensions.

Evolution of Cognitive Radio toward Cognitive Networks is under process, in which Cognitive Wireless Mesh Network (e.g. CogMesh) is considered as one of the enabling candidates aiming at realizing this paradigm change.

Main functions

The main functions of Cognitive Radios are:

- **Spectrum Sensing**: detecting the unused spectrum and sharing it without harmful interference with other users. It is an important requirement of the Cognitive Radio network to sense spectrum holes. Detecting primary users is the most efficient way to detect spectrum holes. Spectrum sensing techniques can be classified into three categories:
  - *Transmitter detection*: cognitive radios must have the capability to determine if a signal from a primary transmitter is locally present in a certain spectrum. There are several approaches proposed:
    - matched filter detection
    - energy detection
    - cyclostationary feature detection
  - *Cooperative detection*: refers to spectrum sensing methods where information from multiple Cognitive radio users are incorporated for primary user detection.
  - *Interference based detection*.

- **Spectrum Management**: Capturing the best available spectrum to meet user communication requirements. Cognitive radios should decide on the best spectrum band to meet the Quality of service requirements over all available spectrum bands, therefore spectrum management functions are required for Cognitive radios. These management functions can be classified as:
  - *spectrum analysis*
  - *spectrum decision*

- **Spectrum Mobility**: is defined as the process when a cognitive radio user exchanges its frequency of operation. Cognitive radio networks target to use the spectrum in a dynamic manner by allowing the radio terminals to operate in the best available frequency band, maintaining seamless communication requirements during the transition to better spectrum.

- **Spectrum Sharing**: providing the fair spectrum scheduling method. One of the major challenges in open spectrum usage is the spectrum sharing. It can be regarded to be similar to generic media access control MAC problems in existing systems

Cognitive radio (CR) versus intelligent antenna (IA)

Intelligent antenna (or smart antenna) is antenna technology using spatial beamforming and spatial coding to cancel interference; however, it requires intelligent multiple or cooperative antenna array. On the other hand, cognitive radio (CR) allows user terminals to sense whether a portion of the spectrum is being used or not, in order to share the spectrum among neighbor users. The following table compares the different points between two advanced approaches for the future wireless systems: Cognitive radio (CR) vs. Intelligent antenna (IA).

Point	Cognitive radio (CR)	Intelligent antenna (IA)
Principal goal	Open Spectrum Sharing	Ambient Spatial Reuse
Interference processing	Avoidance by spectrum sensing	Cancellation by spatial pre/post-coding
Key cost	Spectrum sensing and multi-band RF	Multiple or cooperative antenna arrays
Challenging algorithm	Spectrum management tech	Intelligent spatial beamforming/coding tech
Applied techniques	Cognitive Software Radio	Generalized Dirty-Paper and Wyner-Ziv coding
Basement approach	Orthogonal modulation	Cellular based smaller cell
Competitive technology	Ultra wideband for the higher band utilization	Multi-sectoring (3, 6, 9, so on) for higher spatial reuse
Summary	Cognitive spectrum sharing technology	Intelligent spectrum reuse technology