This document is a collection of research notes compiled by Vipul Naik for MIRI on the distribution of computation in the world. It has not been independently vetted, and is chiefly meant as a resource for other researchers interested in the topic.

1. Answers to major questions on the distribution of computation

Q: *How much of the world's computation is in high-performance computing clusters vs. normal clusters vs. desktop computers vs. other sources?*

A: Computation is split between application-specific integrated circuits (ASICs) and general purpose computing: According to [HilbertLopez] and [HilbertLopez2012], the fraction of computation done by general-purpose computing declined from 40% in 1986 to 3% in 2007. The trend line suggests further decline.

Within general-purpose computing, the split as given on Page 972 (Page 17 of the PDF) in [HilbertLopez2012] for the year 2007 is as follows:

- For installed capacity: 66% PCs (incl. laptops), 25% videogame consoles, 6% mobile phones/PDAs, 3% servers and mainframes, 0.03% supercomputers, 0.3% pocket calculators.
- For effective gross capacity: 52% PCs, 20% videogame consoles, 13% mobile phones/PDAs, 11% servers and mainframes, 4% supercomputers, 0% pocket calculators.

For more detailed data, see Section 2.2 of this document and Section E of [HilbertLopezAppendix].

Q: *What is it being used for, by whom, where?*

A: See the answer above, plus Section 3 of this document.

Q: *How much capacity is added per year?*

A: Growth rates and doubling periods are as follows (based on [HilbertLopez], using data 1986-2007):

- General-purpose computing capacity: growth rate 58% per annum, doubling period 18 months (see Section 2.2 of this document).
- Communication: growth rate 28% per annum, doubling period 34 months (see Section 2.3 of this document).
- Storage: growth rate 23% per annum, doubling period 40 months (see Section 2.4 of this document).
- Application-specific computing: growth rate 83% per annum, doubling time 14 months (see Section 2.2 of this document).

Breakdown of data by time periods is available in [HilbertLopez], and the most important quotes are included in the relevant sections of this document.

Q: *How quickly could capacity be scaled up (in the short or medium term) if demand for computing increased?*

The semiconductor industry is quite responsive to changes in demand, and catches up with book-to-bill ratios as large as 1.4 within 6 months (see Section 3.2 of this document). In addition, the fact that Litecoin, an allegedly ASIC-resistant substitute for Bitcoin, already has ASICs about to be shipped (within two years of launch) also suggests relatively rapid turnaround given large enough economic incentives. In the case of high-frequency trading (HFT), huge investments in nanosecond computing and shaving off milliseconds from Chicago-New York and New York-London cables also suggests quick responsiveness to large incentives.

Q: *How much computation would we expect to be available from custom hardware (FPGA's/ASIC's and their future analogs)?*

A: An increasing fraction (note decline in general-purpose computing's share from 40% in 1986 to 3% in 2007). However, stuff like ASICs and FPGAs can't really be repurposed for other use, so existing ones don't help that much with new tasks.

Q: *What is the state of standard conventions for reporting data on such trends?*

A: The work of Hilbert, Lopez, and others may eventually lead to uniform conventions for reporting and communicating the data, which would allow for a more informed discussion of these. However, Martin Hilbert in particular is skeptical of standardization in the near future, although he believes it possible in principle, see [HilbertStatisticalChallenges] for more. On the other hand, [Dienes] argues for standardization.

2. Different aspects of the extent of computation used

It's worth distinguishing between:

- Capacity, which refers to how much is *available* for use.
- How much is actually used.

Some quick estimates in this direction:

- In his book *The Singularity is Near*, Ray Kurzweil says that about 0.1-1% of capacity on home computers is used.

- [HilbertLopez2012], Page 17 (972 in print version), has an estimate that personal computers use 6-9% of installed capacity.

With this distinction in mind, the following are important aspects of the amount of computation:

- **The number and complexity of processor operations:** From the viewpoint of capacity, this is typically measured in terms of the number of benchmark operations of a particular type per unit time at peak use. Typical examples are MIPS (million instructions per second), FLOPS (floating-point operations per second), and TEPS (traversed edges per second). The fastest supercomputer Tianhe-2 manages 34 petaFLOPS (expected capacity at final deployment ~ 55 petaFLOPS) and the biggest distributed system folding@home manages 18 petaFLOPS. There don't appear to be standardized measures of how much actual computation gets done, but the paper [HilbertLopez] does come up with an estimate of the average number of computations per second that are actually done worldwide (their estimate for 2007 was 6.4 X 10^18 computations per second carried out on general purpose computers, which they say accounted for only 3% of total computation). My current guesstimate for total computation being done would be 0.1-10 zettaFLOPS and current estimate for how much computation can be done if all computers ran at full capacity would be 10-1000 zettaFLOPS (but this would entail prohibitive energy costs and not be sustainable).
- **The amount of disk space used for storage:** Disk space is a crude proxy for the amount of distinct valuable information. Again, we can measure disk space capacity (including unused capacity) or we can measure actual disk space that's been filled. For the latter, we could attempt to measure *unique* disk space, so that we don't double count many different copies of a book or movie that different people have, or we could measure all disk space used with redundancies. [HilbertLopez] estimates that, optimally compressed, humanity had in 2007 a total of 2.9 X 10^20 bytes of information. That's approximately 290 exabytes. The amount was believed to be doubling every 40 months, so assuming that trend continues, we would be close to having about 1 zettabyte of optimally compressed disk space.
- **The amount of communication:** This can again be measured in terms of capacity (the peak capacity for Internet use around the world, measured in bits per second) or in terms of the average bits transferred per second in reality. [HilbertLopez] estimates that in 2007, 2 zettabytes of information were communicated, with a doubling every 34 months, so that we'd be at about 10 zettabytes communicated per year if trends continue.
- **Energy used:** Energy is a major constraint on large-scale computing projects, so a measure of the amount of energy used can help get an idea of how big a project is and what its scaling bottlenecks are. The whole computational ecosystem uses something like 1 petawatt-hour of energy per year. The number is relatively stable, and while growing somewhat, is not doubling any time soon.
- **Physical space used:** The amount of space taken by computing nodes.
- **Physical resources used:** How much of the semiconductor raw materials is being used?
- **Measures of economic size**
- **The sophistication of what's being done**: There's an important sense in which some computations suggest a greater degree of sophistication than others for the same measure of computational resources used. For instance, Google Search and Facebook News Feed generation rely on a very sophisticated AI-like apparatus, whereas streaming large files or Dropbox-style syncing are more mechanical processes even if they use more bandwidth. It's hard to have a reliable measure of sophistication, but this distinction will be important to keep in mind.

2.1. Units of measurement and approximate ballparks

A quick overview of the units we'll be using. Keep in mind that 2^10 = 1024 ~ 10^3 = 1000. There is some notational ambiguity on whether we should take “kilo” to mean 1^10 or 10^3. The distinction can compound and become significant, but our estimates are anyway too uncertain for this to be a critical determinant of precision (explicitly, the ratio 2^70/10^21 is about 1.17, or 17% off from unity – not a very big deal).

2.2. Processor speed

Computational speed (this relates to total amount of computation annually, and is inversely proportional to the amount of time needed to carry out a specific computational task) is measured in FLOPS (floating-point operations per second). Other measures include MIPS (million instructions per second) and TEPS (traversed edges per second) but FLOPS currently seems the most useful measure in terms of useful activity performed for others:

- 1-100 FLOPS: This is sufficient for most single-purpose devices where each activity involves 1 or a few operations of floating-point or lower intensity, such as calculators. A calculator that can do 10 FLOPS will return an answer in 0.1 seconds, which is “good enough” for most humans.
- 1-100 kiloFLOPS: Can carry out small matrix operations and array operations (e.g., updating internally – not necessarily for display – all the entries in a spreadsheet column of ordinary size) in a second. This might be appropriate for scientific calculators (such as those that offer square root, log, and trig functions, and perhaps rudimentary graphing).
- 1-100 megaFLOPS: A basic general-purpose computing device (such as a somewhat outdated laptop or desktop).Can handle displays, video playing, and music playing, though not necessarily very well.
- 1-100 gigaFLOPS: A cutting-edge home laptop or desktop. Can handle simultaneous video, music, downloading, background processing tasks. Slowness becomes visible only for intensive gaming and video and music editing operations in day-to-day use.
- 1-100 teraFLOPS: A server for a company or research lab dedicated to intensive computation. Examples include drug screening, weather prediction, molecular modeling, quantum simulation, movie/video editing (with rapid turnaround time). Other examples include backend processing for mid-sized web companies.
- 1-100 petaFLOPS: The cutting edge for supercomputers these days. Cycle Computing's AWS-based clusters are at the low end (1), BOINC and other large distributed computing are at the middle end (10), and Google, Facebook are probably at the high end (50-100, maybe more). Largest supercomputer Tianhe-2 in China currently 34 petaFLOPS. Enables running large web companies (Google, Facebook) and very rapid weather prediction computations, drug screening and molecular modeling with turnaround times reduced from a week to a few hours.
- 1-100 exaFLOPS: This measures something like the total computational activity of a large country. Some people claim Bitcoin uses that much – note however that Bitcoin computing is not part of general-purpose computing and is done mostly by ASICs. Getting to a supercomputer that can execute at exaFLOPS has sometimes been called “exascale computing” and is something a lot of people want to reach and debate whether we'll ever reach: Would be needed for brain mapping, “instant search” for all the data stored by the NSA, “instantaneous” weather prediction and drug discovery.
- 1-100 zettaFLOPS: Low end here may represent all the *general-purpose* computing that happens in the world today. Higher end may represent all the general-purpose computing that

could happen if all computing equipment currently online were to operate at full capacity, and all the computing (including ASIC-style computing) that happens in the world today.

The distinction between teraFLOPS, petaFLOPS, and exaFLOPS can be thought of as: at teraFLOPS, one just needs to wait for days or weeks on end for the computer to carry out drug discovery-style operations. With petaFLOPS, careful execution can lead one to be done within hours, but the execution might require some general-purpose planning that takes days or weeks (but might still be worth it, given that the per-hour cost of renting petaFLOPS computers (~$2000 for Cycle Computing/AWS) exceeds by a factor of 100X the time cost of programming for optimization.

Once exaFLOPS computing arrives, and assuming it has the same price as petaFLOPS has today (or even, say, 10X more) then what becomes possible is allowing sloppy and experimental code to run the exaFLOPS server, because the feedback is *so* rapid that it's easier to get live feedback than spend days programming. For instance, a search through 205,000 compounds for $33,000 using a Cycle Computing/AWS cluster at Rpeak 1.1 petaFLOPS in November 2013 took 18 hours = 1100 minutes. This required a few days of prior planning and code optimization for the architecture. Now imagine we had exaFLOPS computing at $20,000/hour. Even with code that's 1/10 as efficient, you'd be done in less than 15 minutes, costing $5,000. In this case, it might make more sense to try more experimental code, get the results, learn from them, then run it again, and so on.

See also: http://www.extremetech.com/extreme/122159-what-can-you-do-with-a-supercomputer
http://www.wisegeek.com/what-is-a-teraflop.htm

Best estimates of the total amount of current computational capacity

[HilbertLopez] estimated 6.4 X 10^18 instructions per second in 2007, with a doubling time of 18 months, so that the current estimate (as of 2014) would be about 15-50 times that. This would correspond roughly to the zettaFLOPS range (need MIPS-FLOPS conversion rule discussions).

Best current estimates of the distribution of computational capacity

Computation split between ASICs and general purpose computing: According to [HilbertLopez] and [HilbertLopez2012], the fraction of computation done by general-purpose computing declined from 40% in 1986 to 3% in 2007. The trend line suggests further decline.

Within general-purpose computing, the split as given on Page 972 (Page 17 of the PDF) in [HilbertLopez2012] for the year 2007 is as follows – this is the same data set that they used in [HilbertLopez] (Page 4 of the PDF = Page 62)but with some more explanatory details:

- For installed capacity: 66% PCs (incl. Laptops), 25% videogame consoles, 6% mobile phones/PDAs, 3% servers and mainframes, 0.03% supercomputers, 0.3% pocket calculators.
- For effective gross capacity: 52% PCs, 20% videogame consoles, 13% mobile phones/PDAs, 11% servers and mainframes, 4% supercomputers, 0% pocket calculators.

Best current estimates of growth rates in computational capacity

[HilbertLopez] estimates that, over the period 1986-2007, computational capacity grew as follows:

- General-purpose computational capacity grew at 58% per annum, with a doubling period of 18 months.

- Application-specific computational capacity (which makes up the lion's share of computing) grew at 83% per annum, with a doubling preiod of 14 months.
- Therefore, the relative share of general-purpose computing declined from 40% in 1986 to 3% in 2007.
- The growth rate of general-purpose computing peaked around 1998, at about 80%+ (dotcom bubble times? Large expansion in ownership and use of computers). See Figure 6 of [HilbertLopez]. Value as of 2007 is roughly similar to average 1986-2007 of ~58%.

2.3. Information storage

- A bit can take the values 0 or 1. It is the smallest unit of information storage.
- 1 byte = 8 = 2^3 bits: This is what it takes to store a character in the old ASCII encoding (so one letter of a plain text English document ~ 1 byte)
- 1 kilobyte (KB) = 2^10 ~ 10^3 bytes: This is the most convenient for measuring plain text documents, emails, Facebook posts, text conversations, MIDI-stored music files, low-resolution photos, PDFs of book chapters etc.
- 1 megabyte (MB) = 2^10 ~ 10^3 KB = 2^20 ~ 10^6 bytes: This is most convenient for measuring high-resolution photos, short low-resolution videos, songs (one MP3 minute ~ 0.5-2MB), full-length books or equivalent PDFs. It can be used to measure the size of databases with basic identifying information (without photos, only text) for all people in a city.
- 1 gigabyte (GB) = 2^10 ~ 10^3 MB = 2^30 ~ 10^9 bytes: This can be used to measure sizes for storing movies, large music collections, large book collections (~1000 books), databases of all people in a city along with small thumbnail photos, large high-resolution photo albums, a Wikipedia snapshot (without revision history or photos), very basic information on all people in the world.
- 1 terabyte (TB) = 2^10 ~ 10^3 GB = 2^40 ~ 10^12 bytes: This can be used to measure the space for storing a large movie collection (a huge DVD library equivalent), digital versions of all the books currently in print, all of Wikipedia with photos and revision history, a few days' worth of surveillance footage, a database with photos and biographical information of everybody in the world, the size of the raw footage for a complicated video project.
- 1 petabyte (PB) = 2^10 ~ 10^3 TB = 2^50 ~ 10^15 bytes: This can be used to measure the space taken by all Facebook status updates since the inception of Facebook, the annual disk space taken by videos uploaded to YouTube (76 PB in 2012), the publicly visible text-based part of the World Wide Web (10-100 PB), all the space taken by Facebook photos so far (~10 PB) all the storage space used by Gmail users (though not the storage *capacity* committed, which would get in the exabyte range), the size of the entire video library of Netflix, the size of all the raw footage for a fancy 3D movie.
- 1 exabyte (EB) = 2^10 ~ 10^3 PB = 2^60 ~ 10^18 bytes: This can be used to measure the total amount of storage that a company like Facebook or Google might be planning to build for the next decade keeping in mind continued expansion of service and userbase. It's the order of magnitude for the total capacity committed by Gmail. It might measure the total amount of disk space used by humanity once we exclude redundancies in movie and video storage.
- 1 zettabyte (ZB) = 2^10 ~ 10^3 EB ~ 2^70 ~ 10^21 bytes: This is roughly the total amount of disk space currently used by humanity, and the total disk space capacity would also be measurable in zettabytes.

Best current estimate of total information storage capacity

[HilbertLopez] estimates 290 EB in 2007, with a doubling period of 40 months, so ~1 ZB by now (early 2014) if the trend continues.

Best current estimates of the distribution of information storage

According to [HilbertLopez], information storage was distributed as follows in 2007 (for digital storage): 42% PC hard disks, 21% DVD and Blu-Ray, 11% digital tape, 8% server and mainframe hard disks, 6% CD and mini-disks, 2% portable hard disks, 1% portable media players, and 6% analog video.

Best current estimates of the trends in information storage

According to [HilbertLopez], growth rate 1986-2007 in information storage was 23% per annum, with a doubling period of 40 months. The majority of technological memory has been in digital form since the early 2000s, reaching 94% in 2007. Also, "Storage capacity sloweddown around the year 2000, but accelerated growthhas been occurring in recent years (CAGR of27% for 1986–1993, 18% for 1993–2000, and 26% for 2000–2007) (Table 1)."

2.4. Communication

Let's now consider bitrate measurements. Note that measurements here are customarily reported in bits per second (or kilobits per second, megabits per second) rather than bytes per second, so the corresponding rate in the byte measure would be 1/8 of that (there's also the 2^10 vs 10^3 issue).

- Communication rates of a few (1-10) bits per second: This is useful for nothing else except continued unencrypted communication of status from a small set of options. It might be sufficient for, say, broadcasting expected bus arrival times to stations, or sending signals via remote controllers. General-purpose computing and communication protocols cannot be used.
- Communication rates of a few (0.1-10) kilobits per second: Can load static text-based websites in a few seconds and carry out text-based chatting, but cannot play videos, and downloading a book PDF or e-book or a 5 minute song could take somewhere between several minutes to an hour. Can run Dropbox-style syncing but syncing could take several seconds after every text file update. Mobile phones that can be used for ordinary text and voice communication have a bitrate of 14 kbps.
- Communication rates of a few (0.1-10) megabits per second: Can play streaming video (not very high-definition) and download books in under a minute, but may take on the order of several minutes to an hour to download a movie (basically, the download rate is about the same order of magnitude as the bitrate for actual play). Can use interactive websites that continually update and notify the computer. Dropbox-style syncing is immediate for text and several seconds for large music files.
- Communication rates of a few (0.1-10) gigabits per second: Can download movie-length videos in about a minute. Can sync movies and music files, and upload to YouTube, within seconds. Can even maintain synced offline versions of large websites like Wikipedia, with syncing off by a few minutes.
- Communication rates of a few (0.1-10) terabits per second: Can sync live information from thousands of disparate sources. This is the speed of various components of the Internet backbone.

Best current estimates of the total amount of telecommunications

[HilbertLopez] estimates the humanity communicated 65 EB in 2007 total, which comes to a global bitrate of a few terabits. The telecommunications capacity doubles every 34 months, so it would be about 5x that value by now. However, [HilbertLopez] also noted a faster growth rate in telecomnications capacity in recent years, suggesting a doubling every 18 months, which if extrapolated would imply that the current value is about 10-50x the value in 2007.

Best current estimates of the distribution of telecommunications

[HilbertLopez] estimates that in 2007, 97% of (non-broadcast) telecommunications was fixed Internet, and 1% each for fixed voice phone digital, mobile phone digital data, and mobile phone digital voice. Digital technologies dominated telecommunication, making up 99.9% of telecommunications.

Best current estimates of the trend in telecommunications

Telecommunications capacity grew at 28% per annum in 1986-2007, with a doubling period of 34 months, according to [HilbertLopez]. Further, the growth rate has been accelerating: "The introductionof broadband has led to a continuous accelerationof telecommunication (CAGR of 6% for 1986–1993, 23% for 1993–2000, and 60% for 2000–2007) (Table 1)."

2.5. Timescales (and also, via the speed of light, length scales)

- Seconds: This is a scale of time that humans can experience and calibrate consciously. Delays in seconds are obvious. Computing that takes seconds subjectively “takes time” and does not feel instantaneous.
- Millisecond = 10^(-3) seconds: People do subconsciously adjust to millisecond-scale changes, but generally cannot consciously calibrate these, and things that happen at this scale do “feel instantaneous” to most people. Human reaction time is 100-500 ms (0.1-0.5 s) with exact measure depending on the nature of stimulus, and most humans can't react faster than 150 ms even for the simplest stimuli (see http://www.humanbenchmark.com/tests/reactiontime/). Humans can process auditory and visual stimuli that appear for durations of about 0.1 seconds = 100 ms (lower than the reaction time) – for instance, people who have some musical sense can maintain a rhythm with 8-10 beats per minute without formal training. Some experiments suggest that humans can notice visual stimuli after exposure for as short as 13 ms. The frame rate for human vision is believed to be about 20-25 frames per second, suggesting that humans see discrete frames, one every 40-50 ms. Overall, whatever the minimum threshold at which human sensory perceptions operate, it's likely to be more than 10 ms and highly likely to be more than 1 ms. Millisecond-level improvements are important for people at the backend of designing user-responsive interfaces. Millisecond-level improvements also matter for activities that require computer-to-computer interaction (such as communicating over vast distances) and where improvements beyond the ms level are precluded by the speed of light (for instance, roundtrip speed of light between Chicago and New York is 7.6 ms, current fast cables do it in 13 ms and proposed “Through-air” technology would take 8.5 ms).
- Microsecond = 10^(-6) seconds: These scales are useful for measuring communication between

computers that are part of a cluster in the same or nearby buildings, for instance, a large server facility or a university or big company campus. The speed of light is 3 X 10^8 m/s, so a microsecond describes the time taken by signals to travel 300 meters (~1000 feet). The speed of light constraint can be critical for carrying out communication-intensive distributed computation in a geographically dispersed computing cluster. Microseconds are used to measure performance of trading company servers that are located at the trading exchange.

- Nanosecond = 10^(-9) seconds: This scale is used for computation within a computer or very close-by computing nodes. In a nanosecond, light can travel 0.3 m ~ 1 foot. Talking at this scale is useful when considering the latencies involved for distributing computation within different cores of a processor unit. It also roughly describes the amount of time taken for actual computations in a single core. Note also that a 100 MHz frequency corresponds to one wave for every 10 nanoseconds, so processing of wave signals cannot happen below this scale. Incidentally, here's a video from Grace Hopper about nanoseconds: http://highscalability.com/blog/2012/3/1/grace-hopper-to-programmers-mind-your-nanoseconds.html
- Picosecond = 10^(-12) seconds: Things that happen in picoseconds have to happen over very small distances (3 X 10^(-4) m, which is just barely within the threshold of human perception). We're talking of stuff that's happening within a single processor core. Note that computation here is happening through the movement of electrons, which travel slower than light.
- Femtosecond = 10^(-15) seconds: We're now getting to very small distances: 3 X 10^(-7)m = 300 nm. For comparison, transistor sizes are in the 20-100 nm range.

2.6. Energy numbers

- Kilowatt-hour: Used to measure the energy use of a typical home computer over a year. Values range from 50-200 for laptops with ordinary use and 100-300 for desktops with ordinary use. Mobile phones and tablets are likely under 50.
- Megawatt-hour: Used to measure total annual energy consumption of a typical US household (ranges 5-50) or IT electricity use at a company's in-house server facility.
- Gigawatt-hour: Used to measure total annual energy consumption at a large data center.
- Terawatt-hour: Used to measure total annual energy consumption for a company like Facebook or Google, or total energy consumption of a large First World city.
- Petawatt-hour: Used to measure total annual energy consumption of a large country (~4 PwH for the US), and total computing energy used worldwide (~20 PwH).

Best current estimates of the trend in telecommunications

I'm not aware of any inventorying of total energy use of similar scope and quality to that done by [HilbertLopez] for energy, computation, and communication. Therefore, I need to rely on general information about supply and demand growth.

- Koomey's law is a law for energy similar to Moore's law, and says that the number of computations per unit energy has been doubling every 1.57 years, similar to Moore's law. The total amount of energy expended on computation has thus been growing a lot more slowly than the amount of computation (as noted in Section 2.2, general-purpose computation is doubling every 18 years and application-specific computation is doubling every 14 months, so the growth rates roughly cancel).

- In [KoomeySmartEverything], Jon Koomey argues that so far, the discussion of computation has largely been on the direct resources it uses. However, increasingly, the energy impact of computation will be measured in terms of the resources whose allocation it affects. Computation will become more deeply embedded in our lives and will drive cost savings in resource and energy use. Examples are supply chain management and finance, which control a lot more resources in the real economy relative to the amount of resources they use for computation.
- [KoomeySmartEverything] also argues that application-specific *low-power* computing will become more and more important. This includes devices that consume such low amounts of power that they can operate for several years without the need for recharge or battery change. Some of these will have deep sleep modes that use power in the *picowatt/nonawatt* range, standby modes that use power in the *nanowatt/microwatt* range, and use modes that use power in the *microwatt/milliwatt* range.

2.7. Multipliers

- The number of people in the world is 7 billion (7 X 10^9). Number of Internet connections of various sorts is about 10^9-10^10. Use 10^9 as a multiplier from per capita use of a service or feature to total global usage.
- The number of complex creative works of specific types (books, songs, movies) is generally of the order of 10^5-10^8, with 10^6 a reasonable ballpark. Use 10^6 as a multiplier to go from per item values to total values.

Simulation multiplier (not a standard term, still looking for one): This is a useful concept that tells us whether computers have “arrived” in a particular domain. The question, roughly, is how many seconds of “real” time can be executed in one second of system time. For “playback” type activities, simulation rates of 1 are a breakthrough point where the technology becomes feasible. (Thus, for instance, when it takes 1 second to download 1 second of video, live video streaming becomes, just barely, possible). A 1 mbps bitrate suggests a simulation multiplier of about 5-10 for MP3 video, and of about 1 for YouTube video (and less than one for HD video, making HD streaming impossible). A 20 mbps bitrate (currently the higher end that one can easily obtain reliably) suggests a simulation multiplier of 100-200 for MP3 video, about 20 for YouTube video (so the video can actually buffer very quickly even while playing back) and 5-10 for HD video (depending on the definition and the level of compression). Standard noise removal algorithms on modern home computers take about 1 second per minute of audio processed, suggesting a simulation multiplier of about 60 – which means they can be executed for live streaming audio.

The area where simulation multipliers are still very far below 1 is molecular modeling. In fact, it can take a petaFLOPS supercomputer (a million times as powerful as a home computer) several minutes or even hours to simulate nanoseconds of a molecular model, suggesting a simulation multiplier of about 10^(-12) to 10^(-9). The good news from the molecular modeling perspective is that a lot of the interesting things we need to assess for configuration stability happen in the nanosecond time range.

Similarly, simulating 1 second of human brain activity was reported as having taken a supercomputer with 82,944 processors a total of 40 minutes, suggesting a simulation multiplier of 1/2400. See http://www.extremetech.com/extreme/163051-simulating-1-second-of-human-brain-activity-takes-8294

4-processors

A concept somewhat similar to simulation multiplier (a more parallel version thereof) was considered in [NeumanParkPanek] where they looked at the number of channels being broadcast – in essence, the number of distinct options for live broadcasting that one could choose from at any given time. The number increased from 82 in 1960 to 884 in 2005. Note that this "parallel" measure relies on the bandwidth of the channel whereas the simulation multiplier measure relies more on the speed of processing (though that might be partly parallelizable).

2.8. Trends

Trends (demand-side)

How much can we expect the indicated values to grow over time based on natural growth from the demand-side?

- Bringing more things in existing online categories online: For things that are determined by multiplying per capita values with the population using, we can have gains as more people use. For most global services (such as Facebook, Google, Internet access), we are just one order of magnitude away from population saturation. For "high-level First World Internet use" we may be 2-3 orders of magnitude off (100-1000X). For instance, the typical world Internet user uses less than 1 GB/month, whereas typical Internet users with high-speed connections who get their entertainment, including videos, off the Internet, may rack 25-50 GB/month. So in addition to expanding the Internet to all, we could move people already online to the 50 GB/month mark, meaning a total 500X increase in Internet usage. In addition to looking at population, we could expand the number of *digitized creative works and written material*, but this is likely to be a factor of 10-100X, not a lot more.
- Bringing new forms of stuff online: 3D video, continuous video logging, continuous conversation logging, home automation leading to large amounts of home data stored, 3D scanning/printing designs, etc. These could lead to more disk space used and more computation. With the possible exception of 3D video and continuous video/audio logging, they don't need huge data explosions at the individual level. With 3D video, we are talking of a factor of 10-1000X.
- Large-scale scientific computing, such as molecular dynamics and brain mapping. We don't yet know enough about the optimal storage formats for these, but the explosion in data and computation needs could in principle be huge.

Trends (supply-side)

What are the natural limitations on the rate and ceiling of growth based on whether it's technologically feasible and economically viable for suppliers?

- Supply-side considerations of whether Moore's Law, Kryder's Law, Koomey's Law, Rock's Law, software bloat, etc. will continue. This relates both to whether a particular milestone is technologically feasible, and also whether it will be economically viable to offer products based on that technology.

- Question of how *quickly* suppliers can respond to changes in demand with existing technology or with on-demand technological improvements. For instance, if demand for a particular type of computational device doubled, how quickly would supply catch up? Would prices surge, or stay about the same in the short run? Would prices decline in the long run?

3. Major categories of players

- Companies with a huge web presence, such as Facebook, Google, Amazon, Microsoft, Dropbox, etc.
- Companies that manufacture backend components of computers, including semiconductor manufacturing companies and computer assembly companies.
- Companies that manufacture operating systems and software for desktops, laptops, phones, tablets, etc.
- Governments, such as the US government.
- High-performance computing (HPC) using supercomputer clusters, such as existing supercomputers or temporary instances such as those created by Cycle Computing for scientific and pharmaceutical drug discovery.
- Interactive distributed computing frameworks that may be cooperative or competitive or a mix: Bitcoin (ASIC-heavy), distributed computing for scientific projects, high-frequency trading (ASIC-heavy, low-latency, highly competitive).
- Surreptitious stuff such as botnets.

Obviously, many of these categories have overlap – Google has a huge web presence and also works on Android. Amazon has a web presence of its own and also facilitates HPC by allowing people to rent out large amounts of server space for HPC projects.

Things I haven't looked into – I believe these have huge *destructive* power if they go rogue (i.e., they're going down can throw a wrench in computing), but they have less power to control or command a huge amount of computation at the micro level.

- Companies that manufacture Internet routers and network infrastructure (Cisco, Alcatel, …).
- Companies that own the infrastructure through which stuff is transmitted on the Internet, i.e., telecommunications companies.
- Companies and networks that control the energy grid. Computing is highly reliant on energy.

3.1. Major players in the web space

Organizations include:

- Google: They use a lot of disk space, compute a lot, and use a lot of bandwidth. They're also sophisticated in that they rely on a lot of AI-like algorithms, and are investing aggressively in the next stage of the stuff. Does roughly 0.1-3% of the world's computation and uses a similar proportion of the energy resources devoted to computing.
- Facebook: Similar to Google, maybe about 25-50% the size of Google on most metrics.
- Amazon, in addition to its own website, runs Amazon Web Services (AWS).
- Apple runs iTunes, iCloud, and its App store.

Google

Quantitative measures

Summary: On a wide range of measures, ranging from the amount of computation to the amount of energy used to the amount of disk space used, comes in the range of 0.01-1% of the world and about 0.05-5% of the US (US ~ 20% of the world).

- **The number and complexity of processor operations:** My guesstimate is about 100 petaFLOPS, about 0.01-0.1% of general-purpose computation worldwide, but higher if you look at computation aimed at the outside world. In March 2012, James Pearn wrote a Google+ post guesstimating Google's computational capacity at 40 petaFLOPS, four times the fastest supercomputer at the time (currently, the fastest, Tianhe-2, is capable of 34 petaFLOPS). https://plus.google.com/+JamesPearn/posts/gTFgij36o6u See also earlier estimate in 2008 estimating 20-100 petaFLOPS: http://blogs.broughturner.com/communications/2008/05/google-surpasses-supercomputer-community-unnoticed.html
- **The amount of disk space used for storage (readily accessible index):** The index of the web that Google maintains and uses to answer these queries is about 10-100 PB (roughly the same as the size of the publicly accessible web, a bit less due to compression and a bit more due to indexing for faster search response). Compare to the ~1 ZB estimate of total disk space worldwide, it's 0.001% - 0.01%, however, this is a *rapidly searchable* index of the web.
- **The amount of disk space used for storage (more slowly accessed bulk storage – video):** YouTube is adding about 100 PB of video every year (76 PB in 2012). Total about 0.01-0.1% of the world disk-space-wise. See http://sumanrs.wordpress.com/2012/04/14/youtube-yearly-costs-for-storagenetworking-estimate/ for YouTube estimate.
- **The amount of disk space used for storage (more slowly accessed bulk storage – email):** Gmail has about a billion users. Let's say the average amount of space per user is 10 MB. That totals to 10 PB of disk space occupied by Gmail. The promised space per user is 15 GB, totaling to 15 EB of disk capacity committed by Gmail. So promised space is about 0.1-2% of world disk space, actual space may be a lot less (about 1/1000 of promised space) because most Gmail users are nowhere near close to filling out their space.
- **The amount of communication (web searches):** Google Search processes about 3 billion queries a day (guesstimate). Assuming 30 KB bandwidth communicated per search, that's about 100 TB per day in search traffic costs, or about 36 PB/year. 30 KB is a low estimate because of all the autocompletion and instant search features leading to more continuous data communication. A white paper by Google says it handles in a day is 20 PB (but this includes internal data handling related to searches, not just what's communicated to the world): see http://dl.acm.org/citation.cfm?doid=1327452.1327492 and

http://techcrunch.com/2008/01/09/google-processing-20000-terabytes-a-day-and-growing/

- **The amount of communication (video streaming):** http://sumanrs.wordpress.com/2012/04/14/youtube-yearly-costs-for-storagenetworking-estimate/ estimates that it streams about 16 EB of video a year.
- **The amount of energy used:** Google puts out reports on how much energy it uses per year. The 2012 report (latest at the time of writing) http://www.google.com/green/bigpicture/#/intro/infographics-1 suggests that Google used 3.3 TwH that year, up from 2 TwH in 2010 (see http://gigaom.com/2011/09/08/google-reveals-electricity-use-aims-for-a-third-clean-power-by-2012/). That's about 0.1% of total energy costs in the US, and about 0.5-2% of energy costs devoted to computing. Compared to the world, it's about 0.02% of the world's energy use and 0.3% of the world's energy that goes into computing.

Interest in AI/machine learning

- They've had Peter Norvig and Sebastian Thrun for a while now. Norvig embraces statistical methods to AI and says the goal is to solve specific narrow AI problems rather than get to AGI.
- They hired Andrew Ng, Ray Kurzweil, and Geoffrey Hinton, see e.g. http://www.wired.com/wiredenterprise/2014/01/geoffrey-hinton-deep-learning/ They also started a deep learning research project https://en.wikipedia.org/wiki/Google_Brain where all these people work. Google Research has a section with papers on AI and machine learning: http://research.google.com/pubs/ArtificialIntelligenceandMachineLearning.html
- They recently bought home automation company Nest Labs https://en.wikipedia.org/wiki/Nest_Labs that built a smart thermostat, suggesting ambitions to become a major player with home automation. They spent $3.2 billion.
- They started a Quantum Artificial Intelligence Lab in collaboration with NASA and USRA: https://en.wikipedia.org/wiki/Quantum_Artificial_Intelligence_Lab
- They bought an AI company on January 28, 2014 for $400 million: http://searchenginewatch.com/article/2325629/Google-Buys-AI-Company-DeepMind-May-Have-Big-Plans-for-Search

How critical and secure they are

- **People use Google services as their external memories despite the availability of alternatives for redundancy/independence:** Although Gmail allows users to download their email to a mail client and also to forward email to other accounts for redundancy, most users don't bother doing either, so people can offer be left rudderless when the web service goes down. Similarly, modern browsers offer extensive bookmarking and history capabilities, but people often still go to Google even to navigate to websites they visit regularly. Finally, despite the existence of offline GPS navigators that they can download or use, people still rely on Google Maps to navigate even in places that they've lived in for a while. This suggests a huge *immediate impact* of Google services going down. However, the long run impact would likely be less as people discover substitutes or redundancy measures.
- **Google search is almost never down, but Gmail faces accessibility problems a few times a year**: e.g., they most recently went down January 24, http://techcrunch.com/2014/01/24/gmail-goes-down-across-the-world/
- **Google was hacked by hacker groups believed to be connected with the Chinese**

government in 2009, possibly allowing them to gain access to a lot of sensitive information: See https://en.wikipedia.org/wiki/Operation_Aurora for details. Since then, Google has increased the security of its systems considerably.

Facebook

Quantitative measures

Facebook has approximately a billion users.

- **The number and complexity of processor operations:** Unfortunately, I haven't been able to track any direct data or even any individual's guesstimates and speculation on this. The best bet might be to multiply estimates for Google by the ratio of Facebook's power consumption to Google's (about 0.2).
- **The amount of disk space used for storage (rapidly accessible index):** Facebook claims that it needs over 700 TB of RAM to store all the status updates and comments for all its users (text-based stuff with semantic data). It has implemented a Facebook Graph Search for posts and comments that can execute queries searching and sorting results reading this entire database, and is in the process of rolling this out to users. See https://www.facebook.com/notes/facebook-engineering/under-the-hood-building-posts-search/10151755593228920 for more. Note that this is about 1 or 2 orders of magnitudes less than the index size needed for web search, but queries are semantic and highly personalized, making it challenging in a different way than web search.
- **The amount of disk space used for storage (more slowly accessible bulk storage):** Facebook had about 240 billion photos on its servers as of January 2013, with 350 million new photos being added daily, so about 350 billion by now (January 2014). Total storage was 1.5 PB when they had 100 billion photos, so estimated at about 5 PB now. See http://thenextweb.com/facebook/2013/01/15/facebook-our-1-billion-users-have-uploaded-240-billion-photos-made-1-trillion-connections/ They are looking at “cold storage” solutions for infrequently accessed photos: http://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage/
- **The amount of energy used:** Facebook used 0.7 TWH of energy in 2012, see https://www.facebook.com/green/app_439663542812831 and http://www.datacenterknowledge.com/archives/2013/07/22/facebooks-shifting-power-footprint/

Interest in AI/machine learning

- Facebook is hiring people to work on artificial intelligence problems, mainly with the goal of improving its news feed and suggestions. See http://www.technologyreview.com/news/519411/facebook-launches-advanced-ai-effort-to-find-meaning-in-your-posts/ They recently hired Yann LeCun.

How critical and secure they are

- A lot of our social data is stored on Facebook. Although they do offer ways to download our data, most people don't use it.
- Facebook is not down too often, but there have been outages, usually a few short outages of a few minutes a year in parts of the world.

Amazon

Quantitative measures

- **What little we know:** In addition to server hosting for Amazon.com, Amazon also hosts Amazon Web Services (AWS) intended for people to rent out. Amazon does not release information about AWS the way that Google and Facebook release server information. But guesstimates suggest they have petabytes of data storage, and perhaps exabytes. Amazon does release information on the number of "objects" in its Amazon Simple Storage Service (S3) -- estimated at over 2 trillion, with over 1.1 million requests per second, see http://techcrunch.com/2013/04/18/amazons-s3-now-stores-2-trillion-objects-up-from-1-trillion-last-june-regularly-peaks-at-over-1-1m-requests-per-second/
- **The magnitude of dependence of other companies on Amazon:** Major web companies like Dropbox, Quora, and Reddit use AWS for hosting, suggesting that Amazon has a critical role in the infrastructure (if they decide to go down, they can take down a lot) – in April 2011, Quora and Reddit both went down with Amazon AWS: http://www.eweek.com/c/a/Cloud-Computing/Amazon-EC2-Outage-Disrupts-Service-at-Quora-Reddit-and-Others-136902/. In August 2013, Instagram, Vine, and IFTTT went down due to Amazon outage: http://techcrunch.com/2013/08/25/instagram-vine-and-ifttt-went-dark-thanks-to-amazon-web-services-issues/Even Netflix, a competitor to Amazon, relies on Amazon Web Services infrastructure for backup and redundancy (though not for their main services). See http://aws.amazon.com/solutions/case-studies/netflix/ and http://www.forbes.com/sites/danwoods/2013/01/24/how-netflix-should-recover-from-amazon-addiction/ (Christmas 2012, Netflix downloading got into trouble due to issues with AWS). Dropbox hasn't had any major outage, but that too uses Amazon's Simple Storage Service (S3) to keep data: https://www.dropbox.com/help/7/en

Interest in AI/machine learning

- Amazon does not seem to have explicitly expressed interest in building AI, even though they use a lot of "narrow AI" for their recommendation systems. Amazon's Mechanical Turk is an interesting twist/reversal: see http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html?_r=0

Others that use a lot of bandwidth

These companies use a lot of Internet bandwidth, but don't appear to be doing anything too sophisticated with it:

- Netflix for video downloading
- Dropbox for file syncing
- Some file-sharing services (Megaupload, Bittorrent, etc.) for downloading/uploading
- Skype for voice and video communications

See http://torrentfreak.com/bittorrent-and-netflix-dominate-americas-internet-traffic-111027/

3.2. Major players in the hardware and backend space

Some general remarks on supply side considerations follow. For more information on the economics of the industry, see http://www.semi.org and http://www.investopedia.com/features/industryhandbook/semiconductor.asp

- The industry is heavily cyclical and driven by vagaries in both demand and supply. The general trend seems to be of 1-2 down years followed by 1-2 up years, but estimating the length of down and up periods is hard. The years 2012 and 2013 were down years, but it's believed that there will be resurgence in 2014 and 2015. See for instance http://www.semi.org/en/node/48231 and http://www.semi.org/en/node/48496
- In the long run, it's a *decreasing cost industry*. In other words, in the long run, as demand goes up, prices *fall* as there is sufficient incentive to invest in research and manufacturing equipment that lowers long-run costs. (The short-run story is of course that falling demand causes prices to fall).
- *The industry is quite responsive to increases in demand, suggesting that dramatic increases in computational capacity could be accommodated within a few years.* In cases where the book-to-bill ratio in a given three-month period was quite high (1.4 or higher, meaning that there were 1.4 or more times as many orders as fulfillments, suggesting higher demand than supply), supply 6-9 months later was about the same as demand at the time. For more moderate book-to-bill rations like 1.2, the lag time was 3-6 months. See for instance the data at http://www.semi.org/marketinfo/book-to-bill (I downloaded the Excel file and did some quick calculations based off of that).
- I believe that large-scale consumer demand for technological improvement might be flagging, and this may be a major reason for the lack of significant technological progress in recent years (see also http://www.pcworld.com/article/2030005/why-moores-law-not-mobility-is-killing-the-pc.html). The general claim is that computers are now “good enough” that people aren't making strong demands for further improvements. In the absence of strong demand, the incentive to make significant technological investment is missing.
- One possible route for continued technological growth despite the absence of pressure from the consumer demand side is that niche consumer groups (such as people involved with video editing, animation, gaming, high frequency trading) might provide enough of an impetus to continue investing in research, and that the masses of consumers would then free-ride off the technological improvements. The extent to which this happens depends on (a) how big the niche markets are, (b) how much the technological breakthroughs needed to serve the niche market coincide with technological breakthroughs needed for the general population. For instance, despite the stalling of Moore's law, graphics applications have been steadily improving, and this has spilled over into computers for mainstream people. See the link for the preceding bullet, as well as

http://www.pcworld.com/article/2033671/breaking-moores-law-how-chipmakers-are-pushing-pcs-to-blistering-new-levels.html

IBM

- IBM supplies a lot of server infrastructure to data center. IBM also owns various companies at earlier steps in the supply chain for semiconductor-based manufacturing. Therefore, they can in principle control much of the computation, even if they don't do much themselves.

Interest in AI/machine learning

- IBM developed Watson, an AI that won *Jeopardy!* They're trying to market Watson for use in medical and other niches, but the huge learning time needed to master a domain and the absence of notably superhuman capabilities has discouraged adoption. See http://www.businessweek.com/articles/2014-01-10/ibms-artificial-intelligence-problem-or-why-watson-cant-get-a-job and http://singularityhub.com/2014/01/14/ibm-still-slogging-away-to-market-watsons-ai-smarts-invests-1-billion/

Apple

Apple sells devices but doesn't directly do a lot of computation or communication, with the exception of the iTunes and iCloud. However, they recently seem to have made moves to getting 12 PB storage which suggests ambitions of movie streaming (petabyte-storage can allow one to have a comprehensive movie/video library): http://www.theregister.co.uk/2011/04/06/apple_isilon_order/

3.3. NSA

- **Number and complexity of processor operations:** No reliable estimates here.
- **Amount of disk space used for storage:** The NSA is currently estimated to store a few exabytes of data, see http://www.forbes.com/sites/kashmirhill/2013/07/24/blueprints-of-nsa-data-center-in-utah-suggest-its-storage-capacity-is-less-impressive-than-thought/ This is more than the amount that Facebook and Google hold, mostly because the NSA stores a lot of voice transcripts. But it's more by only one order of magnitude (keep in mind that YouTube adds 76 PB/year, so total may be about half an exabyte, which is just one order of magnitude away). Also, the “few exabytes” figure refers to *capacity*, and the current amount of data actually stored may be an order of magnitude lower.
- **Amount of communication:** Nothing to speak of in terms of direct communication – the NSA relies on secrecy. However, we can estimate their “communication” rate as being essentially the same as the rate to which they're adding to their archive of everything.
- **Amount of energy used:** The total amount of energy used by the National Security Agency for their annual operations seems to be of the same order of magnitude as Google and

Facebook. See https://en.wikipedia.org/wiki/National_Security_Agency#Headquarters for some guesstimates: assuming 100 MW power use on average, that works out to over 0.9 TWH/year.

Interest in AI/machine learning

- The NSA has worked on a number of sophisticated textual analysis techniques that could be classified as narrow AI, see https://en.wikipedia.org/wiki/Mass_surveillance_in_the_United_States

3.4. Brain study initiatives

A number of big projects have been announced to study the brain. If any of them actually take off, they could require a huge amount of computation. But none of them seem poised for takeoff at the moment.

- The US government is funding an initiative to study the brain, see https://en.wikipedia.org/wiki/BRAIN_Initiative. It's been estimated that the initiative could generate 300 EB/year of data. That's two orders of magnitude more than the data storage by any single organization that we're aware of.
- A Blue Brain Project was started in 2005 to study mammalian brains. It is headquartered in Geneva: https://en.wikipedia.org/wiki/Blue_Brain_Project
- The EU has its own Human Brain Project, headquartered in Geneva: https://en.wikipedia.org/wiki/Human_Brain_Project_%28EU%29

3.5. High-performance computing

- Buy versus rent calculation: For projects that need large amounts of high-performance computing for a very short duration, renting works better. For projects that need HPC over a longer duration, buying or building is better.
- **Existing build possibilities – progress in supercomputing:** Until 2012, highest for a supercomputer was < 20 petaFLOPS. In 2013, out came Tianhe-2, operating at 34 petaFLOPS and targeting 55 petaFLOPS eventually at full deployment. See http://www.extremetech.com/tag/supercomputers for articles on supercomputers and https://en.wikipedia.org/wiki/TOP500 for a list of top supercomputers.
- **Rent possibilities:** Cycle Computing (https://en.wikipedia.org/wiki/Cycle_Computing) specializes in providing HPC to clients, typically drug companies and scientific research labs, building on top of Amazon Web Services. Their published examples of usage have steadily improved (September 2011: 30K cores, April 2012: 50K cores, November 2013: 150K cores). The most recent one cost the client $33K over 18 hours, searched 205K compounds, and had a peak capacity (Rpeak) of 1.21 petaFLOPS (compare to Google's guesstimate of 20-100 petaFLOPS or Tianhe-2 supercomputer's value of 34 petaFLOPS, expected to go up to 55 at full deployment). As AWS builds more capacity, expect costs to go down somewhat, even without Moore's law pushing too much.

3.6. Distributed computing projects of various sorts

Bitcoin

- **Number and complexity of processor operations:** Although a lot of computational power around the world is devoted to Bitcoin mining and transactions, this largely uses dedicated mining equipment based on ASICs (application-specific integrated circuits). Therefore, it's not much use taking these resources over for general-purpose computation. In January 2014, hash rate was 13 million Gigahashes/second a week back and 18 million gigahashes/second as of the time of this writing. This claims that global Bitcoin computing power is 64 exaFLOPS or 256x top 500 supercomputers: http://www.forbes.com/sites/reuvencohen/2013/11/28/global-bitcoin-computing-power-now-256-times-faster-than-top-500-supercomputers-combined/ Claim suspicious. Keep in mind, however, that since this computation is happening on ASICs, which account for 97% of computing (compared to general purpose computing that accounts for 3%) it should be measured out of the denominator of all computing rather than out of the denominator of general purpose computing.
- **Amount of energy used:** Assuming 10 W/GH/s energy efficiency would give 1 TWH/year of energy use, comparable with Google and Facebook: http://elidourado.com/blog/bitcoin-carbon/

Litecoin, a close substitute of Bitcoin, was introduced partly with the goal of avoiding a mining arms race. Litecoin uses Scrypt instead of the SHA-1 used by Bitcoin, and Scrypt's mining process is more memory-intensive than processor-intensive, making it unsuitable for ASICs. The hope was that with Litecoin, people would not have an incentive to buy expensive ASIC mining rigs. However, ASICs for Litecoin are on the verge of being introduced, see http://www.coindesk.com/alpha-technology-pre-orders-litecoin-asic-miners/ and more at the Quora question http://www.quora.com/Application-Specific-Integrated-Circuits/Why-is-Bitcoin-believed-to-be-easier-to-game-with-ASIC-than-Litecoin#answers

Note on distributed computing: It's probably true that renting server space comes out cheaper than the electricity costs of a distributed network of home computers. The latter can be cheaper only if you're not paying for electricity. That might happen, for instance, if the users are donating electricity (as with distributed computing for science projects) or if the computers are being used surreptitiously (as happens with botnets).

Distributed computing for scientific projects

- folding@home is estimated to do 18 petaFLOPS. It aims to solve the problem of simulating protein folding.
- The Berkeley Open Infrastructure for Network Computing (BOINC) runs a large number of projects (though not folding@home). Projects include rosetta@home and seti@home. Used 9.2 petaFLOPS in March 2013, current use about 8.3 petaFLOPS. Data available at http://boincstats.com/en/stats/-1/project/detail
- Full list of distributed computing projects, including those not run by BOINC, here: https://en.wikipedia.org/wiki/List_of_distributed_computing_projects

Botnets

- Data on botnets unreliable because of the clandestine mode of operation.
- The webpage https://www.shadowserver.org/wiki/pmwiki.php/Information/Botnets stores information about botnets. Estimate of about 2000 command&control points in January 2014.

High-frequency trading (HFT)

HFT differs somewhat from the rest of the items discussed in that although it's a huge network with a lot of computations, the people involved are competing (intensely) rather than cooperating. So the discussion of HFT is somewhat anomalous.

HFT is closely related to what is called *low latency trading* – trading that relies on very rapid turnaround times. Clearly, a low latency is necessary in order to execute a large number of trades sequentially. However, in principle, low latency trading need not be high-frequency: a trading strategy might involve making a small number of strategic trades as soon as they open up. In practice, however, low latency is demanded largely by people engaged in HFT. They carry out over 55% of trades. Judged by conventional metrics, HFT doesn't carry out a lot of computation. But the computation is carried out and executed quickly, and its effects on financial systems can be huge. So, although I didn't obtain estimates of the computation done by HFT, their true power lies in the effect they have on the global financial system, not the computational resources they use.

Financial markets (including but not necessarily limited to HFT) might be the place where the private return to developing things that are partly in the direction of AGI is highest.

HFTs operate at latencies that approach the theoretical minimum (based on the speed of light). e.g., the New York-Chicago line has a theoretical roundtrip minimum of 7.6 ms, and currently the Spread Networks dark fiber connection takes 13 ms, while other planned air-based lines would take 8.5 ms: http://www.wired.com/business/2012/08/ff_wallstreet_trading/all/

HFT also promotes huge investments in computation with very quick turnaround time. Trades within an exchange are often measured in microseconds, and the trade preparation time is measured in nanoseconds. See http://www.zerohedge.com/news/welcome-sub-nanosecond-markets

Huge cascades of HFT-motivated trades could occur; last example is the 2010 Flash Crash: https://en.wikipedia.org/wiki/2010_Flash_Crash

The company Nanex maintains a ticker tape of HFT transactions and has enabled a lot of people to perform analysis of these transactions: https://en.wikipedia.org/wiki/Nanex The analysis reveals that there are a lot of minor rapid pirce fluctuations that occur at the millisecond level but are over long before humans can even notice them – the 2010 Flash Crash is exceptional in that it persisted long enough for humans to notice. According to http://www.wired.com/wiredscience/2012/02/high-speed-trading/ there were 18,520 such crashes and spikes, between 2006 and 2011.

4. Some additional points

- In the 1980s and 1990s, growth in capacity was driven largely by a growth in the number of devices, i.e., scaling up. In the 2000s, the device count was reaching saturation levels, and growth in capacity was now driven largely by improvements in software and hardware on existing (or replacement) devices.
- Computational capacity on general-purpose computers has been growing about twice as fast as storage and communication (communication increasing slightly faster than storage, but the rate not that different). ASIC computing capacity is growing even faster than general-purpose computing capacity (thrice the rate of communication & storage). See [HilbertSignificance] for the most direct discussion of these points.

References

- [BohnShort] = *How Much Information? 2009 Report on American Consumers* by Roger E. Bohn and James E. Short. Ungated version at http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf
- [Dienes] = *Info Capacity| A Meta Study of 26 "How Much Information" Studies: Sine Qua Nons and Solutions* by István Dienes, available online at http://ijoc.org/index.php/ijoc/article/view/1357
- [HilbertLopez] = *The World's Technological Capacity to Store, Compute, and Communicate Information*, April 2011, by Martin Hilbert and Priscila Lopez –see http://www.uvm.edu/~pdodds/files/papers/others/2011/hilbert2011a.pdf for the direct link and http://www.martinhilbert.net/WorldInfoCapacity.html for the general portal of the authors' research (maintained by Hilbert).
- [HilbertLopez2012] = *How to Measure the World's Capacity to Communicate, Store and Compute Information*, April 2012, by Martin Hilbert and Priscila Lopez http://ijoc.org/ojs/index.php/ijoc/article/view/1562 and http://ijoc.org/ojs/index.php/ijoc/article/view/1563/741
- [HilbertLopezAppendix] = *Methodological and Statistical Background on The World's Technological Capacity to Store, Communicate and Compute Information 2012* by Priscila Lopez and Martin Hilbert, a detailed data supplement for [HilbertLopez] and [HilbertLopez2012]
- [HilbertSignificance] = *How Much Information is There in the "Information Society"?* By Martin Hilbert, *Significance*, 9(4), 8-12. Ungated version at http://www.martinhilbert.net/Hilbert_Significance_pre-publish.pdf
- [HilbertStatisticalChallenges] = *How to Measure "How Much Information"? Theoretical, Methodological, and Statistical Challenges for the Social Sciences* by Martin Hilbert, *International Journal of Communication 6 (2012)*, 1042-1055, available online at http://ijoc.org/index.php/ijoc/article/view/1318/746
- [KoomeySmartEverything] = *Smart Everything: Will Intelligent Systems Reduce Resource Use?* by Jonathan G. Koomey, H. Scott Matthews, and Eric Williams, available online at http://arjournals.annualreviews.org/eprint/wjniAGGzj2i9X7i3kqWx/full/10.1146/annurev-envir

on-021512-110549

- [NeumanParkPanek] = *Tracking the Flow of Information Into the Home: An Empirical Assessment of the Digital Revolution in the US from 1960-2005*. Available online at http://www.wrneuman.com/Flow_of_Information.pdf