Petascale Computational Systems

Gordon Bell and Jim Gray, Microsoft Research
Alex Szalay, The Johns Hopkins University

A balanced cyberinfrastructure is needed to meet growing data-intensive scientific needs.

Most scientific disciplines have long had both empirical and theoretical branches. In the past 50 years, many disciplines—ranging from physics to ecology to linguistics—have also grown a third, computational branch. *Computational science* emerged from the inability to find closed-form solutions for complex mathematical models. Computers make it possible to simulate such models.

In recent years, computational science has evolved to include information management to deal with the flood of data resulting from

- new scientific instruments that, driven by Moore's law, double their data output every year or so;
- the ability to economically store petabytes of data online; and
- the Internet and Grid, which make archived data accessible to anyone, anywhere.

Acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Parallel computers can solve these problems within minutes or hours.

However, the computational efforts of most statistical analysis and data-mining algorithms increase superlinearly. Many tasks involve computing statistics among sets of data points in some metric space. For example, pair algorithms on N points scale as N^2. If the data increases a thousandfold, the work and time can grow by a factor of a million. Many clustering algorithms scale even worse and are infeasible for terabyte-scale datasets.

DATA-CENTRIC COMPUTATION

Next-generation computational systems will generate and analyze petascale information stores. For example, the BaBar detector at the Stanford Linear Accelerator Center currently processes and reprocesses a Pbyte of event data; about 60 percent of the system's hardware budget is for storage and IO bandwidth (www-db.cs.wisc.edu/cidr/papers/P06.pdf).

The Atlas (http://atlasexperiment.org) and CMS (www.cmsinfo.cern.ch) particle detection systems have requirements at least 100 times higher. The Large Synoptic Survey Telescope (www. lsst.org) has needs in the same range: petaoperations per second of processing and tens of Pbytes of storage.

BUILDING BALANCED SYSTEMS

System performance has been improving in line with Moore's law and will continue to do so as multicore processors replace single-processor chips and memory hierarchies evolve. Within five years, a simple, shared-memory multiprocessor will deliver about half a teraoperation per second.

Much of the effort in building Beowulf clusters and supercomputing centers has focused on CPU-intensive TOP500 systems (www.top500.org). Meanwhile, in most sciences the amount of both experimental and simulated data has been increasing even faster than Moore's law because instruments are getting much better and cheaper and storage costs have been decreasing dramatically.

Amdahl's laws

Four decades ago, Gene Amdahl coined many rules of thumb for computer architects:

- *Parallelism*—if a computation has a serial part S and a parallel component P, then the maximum speedup is $S/(S + P)$.
- *Balanced system*—a system needs a bit of I/O per second per instruction per second.
- *Memory*—the Mbyte/MIPS ratio (α) in a balanced system is 1.
- *Input/output*—programs do one I/O per 50,000 instructions.

Although α has increased and caused a slight reduction in I/O density, these "laws" are still generally valid (http://computer.org/proceedings/icde/0506/05060003abs.htm).

In addition, computer systems typically allocate a comparable budget for RAM and for disk storage, which is about 100 times less expensive per Tbyte than RAM. Table 1 captures this 1:100 RAM:disk capacity ratio, along with Amdahl's laws applied to various system powers.

Table 1. Amdahl's laws applied to various system powers.

Operations per second	RAM	Disk I/O bytes/s	Disks for that bandwidth at 100 Mbytes/s/disk	Disk byte capacity (100x RAM)	Disks for that capacity at 1 Tbyte/disk
10^9	Gigabyte	10^8	1	10^{11}	1
10^{12}	Terabyte	10^{11}	1,000	10^{14}	100
10^{15}	Petabyte	10^{14}	1,000,000	10^{17}	100,000
10^{18}	Exabyte	10^{17}	1,000,000,000	10^{20}	100,000,000

Scaled to a petaoperations-per-second machine, Amdahl's laws imply the need for

- parallel software to use that processor array and a million disks in parallel;
- a Pbyte of RAM;
- 100 Tbytes/s of I/O bandwidth and an I/O fabric to support it;
- one million disk devices to deliver that bandwidth (at 100 Mbytes/s/disk); and
- 100,000 disks storing 100 Pbytes of data produced and consumed (at 1 Tbyte/disk), which is 10 times fewer than the number of disks required by the bandwidth requirement.

A million disks to support a petascale processor's IO needs is a daunting number. If a petascale system is configured with fewer disks, the processors will probably spend most of their time waiting for IO and memory—as is often the case today.

Petascale systems

There are precedents for such petascale distributed systems at Google, Yahoo!, AOL, and MSN (http://doi.acm.org/10.1145/945450). These systems have tens of thousands of processing nodes (approximating a petaoperation per second) and have about 100,000 locally attached disks to deliver the requisite bandwidth. Although they aren't commodity systems, they're in everyday use in many data centers.

Once empirical or simulation data is captured, huge computational resources are needed to analyze the data and visualize the results. Analysis tasks involving Pbytes of information require petascale storage and I/O bandwidth. In addition, the data must be reprocessed each time a new algorithm emerges or researchers pose a fundamental new question, generating even more I/O.

More importantly, to be useful, these databases require the ability to process information at a semantic level. The data must be curated with metadata, stored under a schema with a controlled vocabulary, and organized for quick and efficient temporal, spatial, and associative search. Petascale database systems will be a major part of any successful petascale computational facility and require substantial software investment.

DATA LOCALITY

Moving a byte of data across the Internet has a well-defined cost (http://doi.acm.org/10.1145/945450). Moving data to a remote computing facility is worthwhile only if performing the analysis requires more than 100,000 CPU cycles per byte of data. SETI@home, cryptography, and signal processing have such CPU-intensive profiles. However, most scientific tasks are more in line with Amdahl's laws and much more information-intensive, with CPU:IO ratios well below 10,000:1.

For less CPU-intensive tasks, colocating the computation with the data is preferable. In a data-intensive world where Pbytes are common, however, it's important to colocate computing power with the databases rather than moving the data across the Internet to a "free" CPU. If the data must be moved, it makes sense to store a copy at the destination for later reuse. Managing this data movement and caching poses a substantial software challenge. Much current middleware assumes that data movement is free and discards copied data after use.

COMPUTATIONAL PROBLEM SIZES

Scientific computation task sizes depend on the product of many independent factors. Quantities formed as a product of independent random variables follow a lognormal distribution (E.W. Montroll and M.F. Shlesinger, "Maximum Entropy Formalism, Fractals, Scaling Phenomena, and 1/*f* Noise: A Tale of Tails," *J. Statistical Physics*, vol. 32, no. 2, 1983, pp. 209-230). As a result, the sizes of scientific computational problems obey a power law wherein the problem size and the number of such problems are inversely proportional—there are a small number of huge jobs and a huge number of small jobs.

This situation is quite evident in US computing today. Thirty years ago, supercomputers were the mainstay of computational science. However, today's four-tier architecture—including tier-1 supercomputers, tier-2 regional centers, tier-3 departmental Beowulf clusters, and tier-4 single workstations—reflects the problem-size power law.

BALANCED CYBERINFRASTRUCTURE

What's the best allocation of cyberinfrastructure investments in light of Amdahl's laws, the problem-size power law, and the move to data-centric science? There certainly must be two high-end tier-1 international data centers serving each discipline that

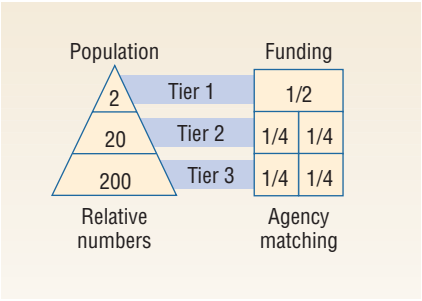

Figure 1. Balanced cyberinfrastructure. Government should fund tier-1 centers and half of tier-2 and tier-3 centers on a cost-sharing basis.

- allow competition,
- encourage design diversity, and
- leapfrog each other every two years.

These tier-1 facilities should contain much of science's huge archival datasets and can only be built as a national or international priority.

What should US government agencies and industry do about the other tiers? They could make funding the tier-2 and tier-3 systems entirely the universities' responsibility, but that would be a mistake.

We believe that available resources should be allocated to benefit the broadest cross-section of the scientific community. Given the power-law distribution of problem sizes, this means that about half of funding agency resources should be spent on tier-1 centers at the petascale level and the other half dedicated to tier-2 and tier-3 centers on a cost-sharing basis, as Figure 1 shows.

One of the most data-intensive science projects to date, the CERN Large Hadron Collider (http://lhc.web.cern.ch/lhc), has adopted exactly such a multitiered architecture. The hierarchy of an increasing number of tier-2 and tier-3 analysis facilities provides impedance matching between the individual scientists and the huge tier-1 data archives. At the same time, the tier-2 and tier-3 nodes provide complete replication of tier-1 datasets.

EXAMPLE TIER-2 NODE

Most funding for tier-2 and tier-3 centers today splits costs between the federal government and the host institution. It's difficult for universities to obtain private donations for computing resources because they depreciate so quickly. Donors generally prefer to give money for buildings or endowed positions, which have a long-term staying value. Government funding is therefore crucial for tier-2 and tier-3 centers in a cost-sharing arrangement with the hosting institution.

For example, The Johns Hopkins University received a National Science Foundation grant toward the computers for a tier-2 center it's building. JHU matched the NSF funds 125 percent to provide the hosting facility and staff to run it. Other institutions have had similar experiences setting up large computing facilities. The price of computers is less than half the cost, and universities can meet those infrastructure demands only if federal agencies seed the tier-2 and tier-3 centers.

Placing all the financial resources at one end of the power-law distribution would create an unnatural infrastructure incapable of meeting the increasingly data-centric requirements of most midscale scientific experiments. At the system level, focusing on CPU harvesting would also create an imbalance. Funding agencies should support balanced systems, not just CPU farms, as well as petascale IO and networking. They should also allocate resources for a balanced tier-1 through tier-3 cyberinfrastructure.

Gordon Bell *is a senior researcher in the Media Presence Research Group at Microsoft's Bay Area Research Center (BARC). Contact him at GBell@Microsoft.com.*

Jim Gray *is a distinguished engineer in the Scalable Servers Research Group at BARC. Contact him at Gray@Microsoft.com.*

Alex Szalay *is a professor in the Department of Physics and Astronomy at The Johns Hopkins University. Contact him at Szalay@jhu.edu.*

Editor: Simon S.Y. Shim, Department of Computer Engineering, San Jose State University; sishim@email.sjsu.edu