Petascale Computational Systems: Balanced CyberInfrastructure in a Data-Centric World

Gordon Bell[1], Jim Gray[1] and Alex Szalay[2]
1. Microsoft Research
2. The Johns Hopkins University
GBell@Microsoft.com, Gray@Microsoft.com, Szalay@jhu.edu
September 2005

Abstract: Computational science is changing to be *data intensive*. Super-Computers must be *balanced system,* not just CPU farms but also petascale IO and networking arrays. Anyone building CyberInfrastructure should allocate resources to support a balanced *multi-Tier* design – from a few huge datacenters to many university-scale systems.

Computational Science and Data Exploration

Computational Science is a new branch of most disciplines. A thousand years ago, science was primarily *empirical*. Over the last 500 years each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding. Today most disciplines have both empirical and theoretical branches. In the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical and computational biology, or physics, or linguistics). Computational Science has meant simulation. It grew out of our inability to find closed form solutions for complex mathematical models. Computers can simulate these complex models.

Over the last few years Computational Science has been evolving to include information management. Scientists are faced with mountains of data that stem from four trends: (1) the flood of data from new scientific instruments driven by Moore's Law – doubling their data output every year or so; (2) the flood of data from simulations; (3) the ability to economically store petabytes of data online; and (4) the Internet and computational Grid that makes all these archives accessible to anyone anywhere, allowing the replication, creation, and recreation of more data [2].

Acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. By using parallelism, these problems can be solved within fixed times (minutes or hours). In contrast, most statistical analysis and data mining algorithms are nonlinear. Many tasks involve computing statistics among sets of data points in some metric space. Pair-algorithms on N points scale as N^2. If the data increases a thousand fold, the work and time can grow by a factor of a million. Many clustering algorithms scale even worse. These algorithms are infeasible for terabyte-scale datasets.

Computational Problems are Becoming Data-Centric

Next generation computational systems and science instruments will generate petascale information stores. The computation systems will often be used to analyze these huge information stores. For example BaBar processes and reprocesses a petabyte of event data today. About 60% of the BaBar hardware budget is for storage and IO bandwidth [1]. The Atlas and CMS systems will have requirements at least 100x higher (http://atlasinfo.cern.ch, http://cmsinfo.cern.ch/). The Large Synoptic Survey Telescope (LSST) has requirements in the same range: peta-ops of processing and tens of petabytes of storage (http://lsst.org/).

Amdahl's Laws – Building Balanced Systems

System performance has been improving with Moore's law and it will continue as multi-core processors replace single processor chips and as memory hierarchies evolve. Within five years, we expect a simple, shared memory multiprocessor to deliver about ½ tera-ops. Much of the effort in building Beowulf clusters and the supercomputing centers has been focused on CPU-intensive TOP-500 rankings (http://Top500.org/) and has ignored IO metrics. Meanwhile, in most sciences the amount of data (both experimental and

simulated) has been increasing even faster than Moore's law because the instruments are getting so much better and cheaper, and because storage costs have been improving much faster than Moore's law.

Gene Amdahl coined many rules of thumb for computer architects. Surprisingly, 40 years later, the rules still apply [4]:

Amdahl's parallelism law: *If a computation has a serial part S and a parallel component P, then the maximum speedup is S/(S+P).*
Amdahl's balanced system law: *A system needs a bit of IO per second per instruction per second.*
Amdahl's memory law: *α=1: that is the MB/MIPS ratio (called alpha (α)), in a balanced system is 1.*
Amdahl's IO law: *Programs do one IO per 50,000 instructions*

In [4], it is shown that α has increased and that has caused a slight reduction in IO density, but these "laws" are still decent rules-of-thumb. In addition to Amdahl's laws, computer systems typically allocate comparable budgets for RAM and for disk. A terabyte of disk storage is about one hundred times less expensive than RAM. This 1:100 RAM:Disk capacity ratio and the Amdahl laws are captured in the following spreadsheet:

Table 1. Amdahl's Balanced System's Laws Applied to Various System Powers.

OPS	OPS	RAM	Disk IO Byte/s	Disks for that Bandwidth @ 100MB/s/disk	Disk Byte Capacity (100x RAM)	Disks for that Capacity @ 1TB/disk
giga	1E+09	gigabyte	1E+08	1	1E+11	1
tera	1E+12	terabyte	1E+11	1,000	1E+14	100
peta	1E+15	petabyte	1E+14	1,000,000	1E+17	100,000
exa	1E+18	exabyte	1E+17	1,000,000,000	1E+20	100,000,000

Scaled to a peta-operations-per-second machine, these rules imply

- There must be parallel software to use that processor array and a million disks in parallel,
- The system will have a petabyte of RAM , and
- 100 terabytes/sec of IO bandwidth and an IO fabric to support it, and
- 1,000,000 disk devices to deliver that bandwidth (at 100 MB/s/disk), and at least
- 100,000 disks storing 100 PB of data produced and consumed by this peta-ops machine (at 1TB/disk) (note that this is 10x fewer than the number of disks required by the bandwidth requirement.)

The storage bandwidth number is daunting – a million disks to support the IO needs of a peta-scale processor. 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. There are precedents for such petabyte-scale distributed systems at Google, Yahoo!, and MSN Search [5]. Those systems have tens of thousands of processing nodes (approximately a peta-ops) and have ~ 100,000 locally attached disks to deliver the requisite bandwidth. These are not commodity systems, but they are in everyday use in many datacenters.

Once empirical or simulation data is captured, huge computational resources are needed to analyze the data and huge resources are needed to visualize the results. Analysis tasks, involving petabytes of information require petascale storage and petascale IO bandwidth. Of course, the data needs to be reprocessed each time a new algorithm is developed and each time someone asks a fundamentally new question. That generates even more IO.

Even more importantly, to be useful, these databases require an ability to process the information at a semantic level – rather than just being a collection of bytes. The data needs to be curated with metadata, stored under a schema with a controlled vocabulary, and need to be indexed and organized for quick and efficient temporal, spatial, and associative search. These peta-scale database systems will be a major part of any successful petascale computational facility and will require substantial software investment.

Data Locality—Bringing the Analysis to the Data

There is a well-defined cost associated with moving a byte of data across the Internet [3]. It is only worth moving the data to a remote computing facility if the problem requires more than 100,000 CPU cycles per byte of data. SETI@home, cryptography, signal processing are examples that have such CPU intensive profiles; but most scientific tasks are more in line with Amdahl's laws and are much more information intensive having CPU:IO ratios well below 10,000:1.

In a data-intensive world, where petabytes are common it is important to co-locate computing power with the databases rather than planning to move 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 it is used.

Computational Problem Sizes Follow a Power Law

The sizes of scientific computations depend on the product of many independent factors. Quantities formed as a product of independent random variables follow a lognormal distribution [6]. As a result, the sizes of scientific computational problems obey a power law where 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. One can see this quite well in the current situation in US computing. Thirty years ago, supercomputers were the mainstay of computational science. Today the whole range is filled from a few Tier-1 supercomputers, to Tier-2 regional centers, to huge numbers of Tier-3 departmental Beowulf clusters, and vast numbers of Tier-4 workstations. This 4-tier architecture reflects the problem size power-law.

Building a Balanced CyberInfrastructure

What is 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 must certainly be two high-end Tier-1 international data centers serving each discipline that (1) allow competition, (2) allow design diversity, and (3) that leapfrog one another every two years. The data centers should have much of science's huge archival datasets. The Tier-1 facilities can only be built as national or international priorities. But, what should 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 the 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 part of funding agency resources should be spent on national, high-end Tier-1 centers at the petaop level; and comparable amounts (about 50%) should be allocated to co-fund Tier-2 and Tier-3 centers. The resulting division would be a balanced allocation of resources with government agencies funding about ½ the Tier-2 and Tier-3 centers, while institutions and other mission-agencies would fund the other ½ of these lower-tiers on a cost-sharing basis.

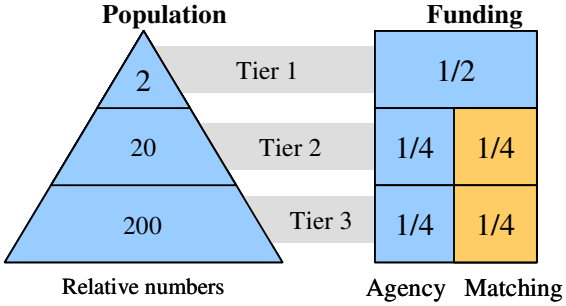

Figure 1: A balanced tiered system allocates approximately equal resources to each tier. The next lower tier is ten times smaller and more numerous. Governments should fund Tier-1 and half of the lower tiers on a cost-sharing basis.

It is interesting to note, that one of the most data intensive science projects to date, the CERN Large Hadron Collider, has adopted exactly such a multi-tiered 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 the Tier-1 datasets.

An Example Tier-2 Node: Balanced Architecture and Cost Sharing

Most Tier-2 and Tier-3 centers today split costs between the government and the host institution. It is difficult for Universities to get private donations towards computing resources because they depreciate so quickly. Donors generally prefer to donate money for buildings or endowed positions, which have a long term staying value. Therefore, government funding is crucial for Tier-2 and Tier-3 centers on a cost-sharing arrangement with the hosting institution.

To give a specific example, The Johns Hopkins University (JHU) is building a Tier-2 center. It received an NSF MRI grant towards the computers for this facility. Incremental power and cooling needs pushed the building above the original design threshold requiring an upgrade to the hosting facilities. In addition, a 50% FTE systems person runs the facility. As a result, JHU's matching funds were 125%. Other institutions have had similar experiences when setting up larger computing facilities: the price of computers is less than half the cost and the Universities are willing to provide those infrastructure costs if NSF seeds the Tier-2 and Tier-3 centers.

Summary

In summary we would like to emphasize the importance of building balanced systems, reflecting the needs of today's science, and also of building a balanced cyberinfrastructure. Placing all the financial resources at one end of the power-law (or lognormal) distribution would create an unnatural infrastructure, where the computing needs of most mid-scale scientific experiments will not be met. On the system level, placing all the focus on CPU harvesting will also tip the balance.

In this short article we argued:

1. Computational science is *changing* to be data intensive.
2. Funding agencies should support *balanced systems*, not just CPU farms but also petascale IO and networking.
3. They should allocate resources to support a balanced *Tier-1 through Tier-3 cyberinfrastructure*.

References

[1] Becla, J., Daniel L. Wang, D.L., "Lessons Learned from Managing a Petabyte," CIDR 2005: Asilomar 5-7 Jan. 2005, pp. 70-83, http://www-db.cs.wisc.edu/cidr/papers/P06.pdf

[2] *Getting Up To Speed, The Future of Supercomputing,* Graham, S.L. Snir, M., Patterson, C.A., (eds), NAE Press, 2004, ISBN 0-309-09502-6

[3] Gray, J., "Distributed Computing Economics", *Computer Systems Theory, Technology, and Applications, A Tribute to Roger Needham*, A. Herbert and K. Sparck Jones eds., Springer, 2004, pp 93-101, also MSR-TR-2003-24, March 2003, http://research.microsoft.com/research/pubs/view.aspx?tr_id=655

[4] Gray, J., Shanoy, P.J., "Rules of Thumb in Data Engineering," Proc ICDE200, San Diego, March 1-4, 2000. IEEE Press., pp 3-12, http://computer.org/proceedings/icde/0506/05060003abs.htm

[5] Ghemawat, S., Gobioff, H., Leung, S., "The Google File System," Proc. SOSP 2003, Lake George, New York, October 19-22, 2003, ACM Press, pp. 29-43, http://doi.acm.org/10.1145/945450

[6] Montroll, E.W. and Shlesinger, M.F.: "Maximum entropy formalism, fractals, scaling phenomena, and 1/f noise: a tale of tails," J. *of Statistical Physics*, Vol. 32, pp. 209-230, (1983).