A Call to Arms: A New Manifesto for Systems

David Patterson, University of California at Berkeley, Marh 31, 2002

Let's start by congratulating the systems community for a 10,000 times performance increase over the last 20 years; what took a year to calculate in 1982 takes less than an hour in 2002. As a result, astronomers today explore the universe inside a computer cluster: $10,000 today provide a scientists with terabytes of storage, gigabytes of memory, and processors running at multiple gigaherz. Thus, most scientists today are computational scientists, and all future scientists should train in our field. For the rest of us, free services like Google put billions of information sources at our fingertips, transforming vague remembrances into total recall, thereby amplifying our knowledge.

Alas, there are unintended consequences of our obsession with performance. First, we have developed a brittle technology that is frustrating to use. Doctors and lawyers are installing device drivers, system administrators are spending their nights battling bugs in expensive software, and user frustration even leads to violence.[1] Second, our computers--from gadgets to Internet services--are undependable, with newspaper stories about outages becoming cliches. Third, computer security is an oxymoron, and our privacy has probably never been more public. Finally, advances in cost-performance have made computers ubiquitous and so the world now depends on them, despite the fact that we never demonstrated nor promised that we merit that trust.

My nightmare is that in twenty years computers will be another 10,000 times faster, but they will be just as frustrating and no more secure or dependable than they are today.

A New Manifesto for Systems

It is time to broaden the systems agenda. It may even be even time to invent a new foundation on which to build the next generation of information technology. It is much easier to complain than to propose, so here are three goals for a new systems manifesto:

1. **Synergy with humanity**: We need to make the technology matches human abilities, both for users of gadgets and for those who operate the services accessed by those gadgets. Thus, we must care as much about the human cost of operating information technology as about the cost of purchasing it.
2. **Dependability**: We need to create a technology that the world can really depend upon, since its already doing it with a technology that doesn't deserve our trust.
3. **Security/Privacy**: Let's not be the enablers of Big Brother. We need to invent the technologies that helps make our societies secure without sacrificing privacy.

Technology Push

The three goals above pull us towards a new agenda, but there are also technological opportunities that push towards innovation.

Architects and compiler writers have achieved the four orders of magnitude speedup in large part by taking advantage of the implicit parallelism. Designers today believe we have exploited almost all the available implicit parallelism, so even to *maintain* past rates of performance improvement, we will likely have to change. A simple pipelined RISC processor takes less than 0.5 mm^2, so hundreds can fit on a chip today

[1] A Mori survey in Britain found that more than 12% have seen their colleagues bully the IT department when things go wrong, while 25% of under 25 year olds have seen peers kicking their computers. Some 2% claimed to have actually hit the person next to them in their frustration. HCI Prof. Helen Petrie says, " it starts in the mind, then becomes physical, with shaking, eyes dilating, sweating, and increased heart rate. You are preparing to have a fight, with no one to fight against." From *Net effect of computer rage*, by Mark Hughes-Morgan, AP, 2/25/02.

and thousands will in the future. Putting that into perspective, the first microprocessor used only 2300 transistors. How should we design and program computers when the basic building block is a processor rather than a transistor?

Second, we are living in a time of wretched excess of hardware resources, even if we ignore the processor. For example, 20 years ago a gigabyte of DRAM memory cost $10,000,000, but today it is almost $100. Similarly, a gigabyte of disk storage cost $100,000 in 1982, but now it is dropping to $1. We may soon have ten gigabytes in a single memory module and a terabyte in a single disk!

We appear to have a wealth of resources to push a new manifesto.

Inspiring Progress

What we can't measure, we can't improve. Hence, we need to create benchmarks to measure progress towards our new goals. In addition to the obvious benefits, quantitative measurements facilitate publications by helping reviewers quantify the value of a set of ideas. They also comfort systems researchers used to quantitative performance assessment whom may be apprehensive about changing directions.

These benchmarks will involve people, as they are the critical to all these new goals: people are the weak link in dependability[2] and security/privacy[3] as well as in human-computer synergy. Although rare in systems research, human subject experimenta are the core of the social science research. Hence, we may need to learn new research skills to embrace our new manifesto.

Once we can measure it, one way to encourage progress in a new direction is to reward it. For example, a jury awards Gordon Bell Prizes each year in high performance computing[4]. The jury ensures that performance is for real problems, and gives awards in multiple categories. The database community gives similar awards for the fastest sort[5]. If a community or a benefactor awarded prizes for advances towards our proposed three goals, we can both inspire and document rapid progress on our new manifesto.

A Call to Arms

This paper is a call to arms for systems researchers: we *must* broaden our research agenda beyond performance, so that our legacy does not remain as fast but flaky.

Perhaps the current stack of technology, which has been our foundation for the last 20 years, can evolve to address the issues above. Perhaps instead we must take the road less traveled, to start afresh with new architectures, new software systems, new programming systems, and new applications.

No matter which road we take, the result must enhance the human experience, offer real security and privacy, and be so dependable that the world can count on it. Such an advance in the next 20 years may be more impressive than the performance improvement of the last 20 years.

---

[2] More than half the outages in 2001 for the public switched telephone system and for three Intent sites were due to operator error. See *Recovery-Oriented Computing (ROC): Motivation, Definition, Techniques, and Case Studies*, by D. Patterson *et al*, UC Berkeley CSTR UCB//CSD-02-1175, March 15, 2002.

[3] Less than half of novice users could properly set up a secure email transmission using a security system known to be easy to use. See *Why Johnny Can t Encrypt: A Usability Evaluation of PGP 5.0*. by Alma Whitten and J.D. Tygar. In Proceedings of the 9th USENIX Security Symposium, August 1999.

[4] The annual Gordon Bell Prizes are awarded to recognize achievements in high-performance computing, with Gordon Bell donating the $5000 prize. A purpose of the award is to track the progress over time of parallel computing. Awards are in peak performance and price/performance in a genuine applications.

[5] Coordiinted by Jim Gray, the awards are for how much to sort for a penny, how much to sort in a minute, and how long to sort a terabyte. See http://www.research.microsoft.com/barc/SortBenchmark/

DAVID A. PATTERSON (University of California at Berkeley) has taught computer architecture since joining the faculty in 1977, and is holder of the E.H. and M.E. Pardee Chair of Computer Science.

At Berkeley, he led the design and implementation of RISC I, likely the first VLSI Reduced Instruction Set Computer. This research became the foundation of the SPARC architecture, currently used by Fujitsu, Sun Microsystems, and others. (In 1996 Microprocessor Report and COMDEX named SPARC as one of the most significant microprocessors as part of the celebration of the 25th anniversary of the microprocessor.) He was also a leader of the Redundant Arrays of Inexpensive Disks (RAID) project, which led to reliable storage systems from many companies. These projects led to three distinguished dissertation awards from the Association for Computing Machinery (ACM). He was also involved in the Network of Workstations (NOW) project, which led to cluster technology used by Internet companies such as Inktomi. His current research interests are in building novel microprocessors using Intelligent DRAM (IRAM) for use in portable multimedia devices, and in Recovery Oriented Computing (ROC) to provide computers for Internet services that are highly available, easily maintained, and gracefully evolve.

In the past he has been chair of the CS Division in the EECS department at Berkeley, the ACM Special Interest Group in Computer Architecture (SIGARCH), and the Computing Research Association (CRA). He is currently a member of the National Academy of Sciences Computer Science and Telecommunications Board and the CRA Board. He has consulted for several companies, including Digital (now Compaq), Hewlett Packard, Intel, and Sun Microsystems, and is on the advisory board of several startup companies. He is also co-author of five books, including two with John Hennessy, President of Stanford University.

He is a member of the National Academy of Engineering, is a Fellow of the Computer Society of the Institute of Electrical and Electronic Engineers (IEEE), and is also a Fellow of the ACM. He received the inaugural Outstanding Alumnus Award of the UCLA Computer Science Department as part of its 25th Anniversary. In 1995 he received the IEEE Technical Achievement Award. In 1998 he shared the inaugural Test of Time Award with Garth Gibson and Randy Katz, given by the Special Interest Group on Management of Data (SIGMOD) to the most influential paper from the SIGMOD proceedings 10 years earlier. The following year they also shared the IEEE Reynold B. Johnson Information Storage Award "for the development of Redundant Arrays of Inexpensive Disks (RAID)." In 2000 he shared the IEEE John von Neumann Medal with John Hennessy "for creating a revolution in computer architecture through their exploration, popularization, and commercialization of architectural innovations."

His teaching has been honored by his department in 1998 with the Diane S. McEntyre Award for Excellence in Teaching, by the University of California in 1982 with the Distinguished Teaching Award, by the ACM in 1991 with the Karl V. Karlstrom Outstanding Educator Award, by IEEE in 1996 with the Undergraduate Teaching Award, and by the IEEE again in 2000 with the James H. Mulligan, Jr. Education Medal "for inspirational teaching through the development of creative curricula and teaching methodology, for important textbooks, and for effective integration of education and research missions."