Presentation to 2004 Workshop on Extreme Supercomputing Panel:

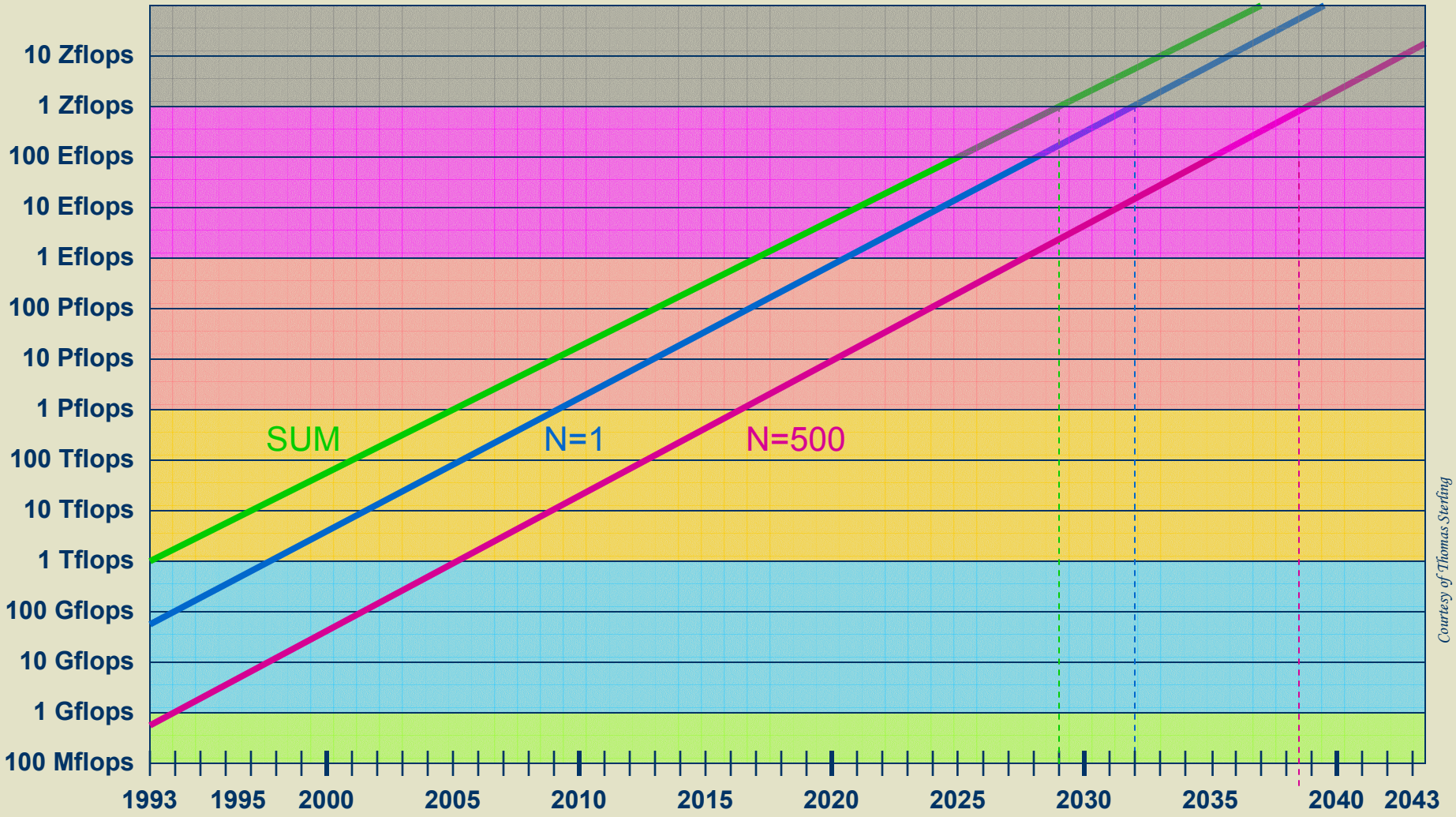
Roadmap and Change

How Much and How Fast

Thomas Sterling
California Institute of Technology
and
NASA Jet Propulsion Laboratory
October 12, 2004

2^9 Years Ago Today

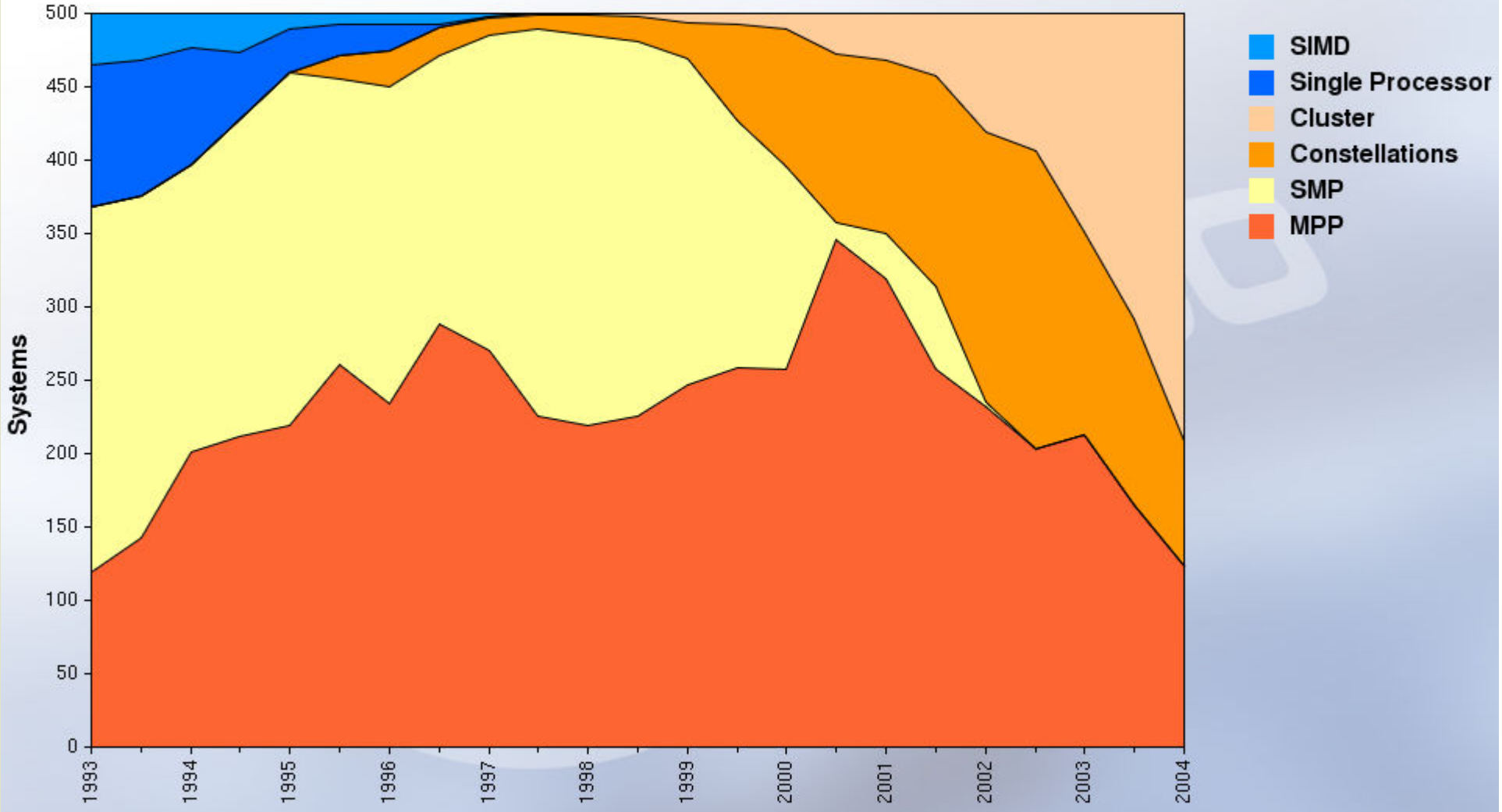
Linpack Zettaflops in 2032



Courtesy of Thomas Sterling

TOP500 SUPERCOMPUTER SITES

Architectures / Systems

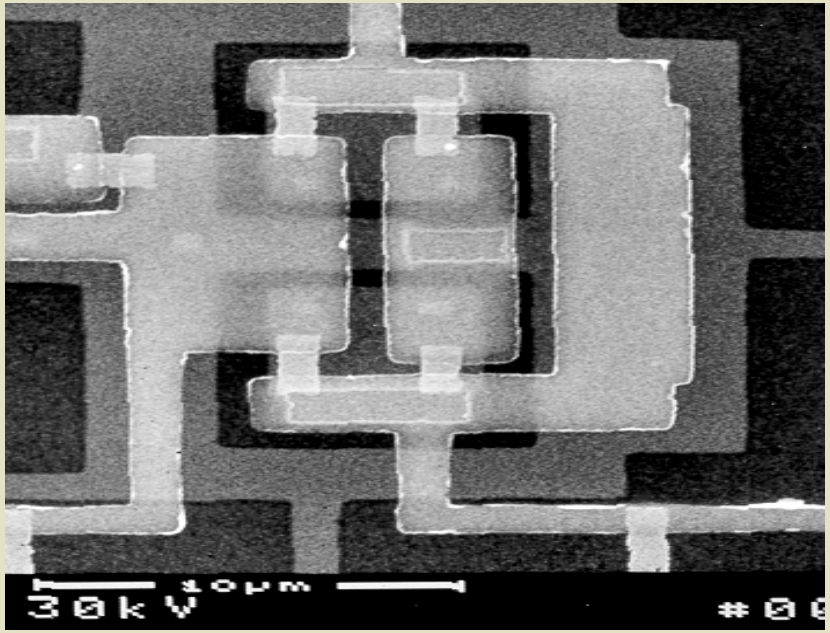


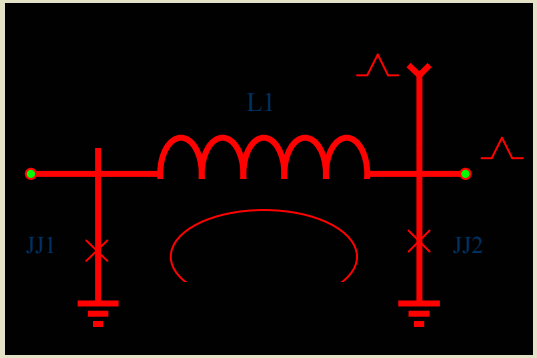
The Way We Were: 1974

- IBM 370 market mainstream
  - Approx. 1 Mflops
- DEC PDP-11 geeks delight
- Seymour Cray started working on Cray-1
  - Approx. 100 Mflops
- 2nd generation microprocessor
  - e.g. Intel 8008
- Core memory
- 1103 1Kx1 DRAM chips
- Punch cards, paper tapes, teletypes, selectrics

What Will Be Different

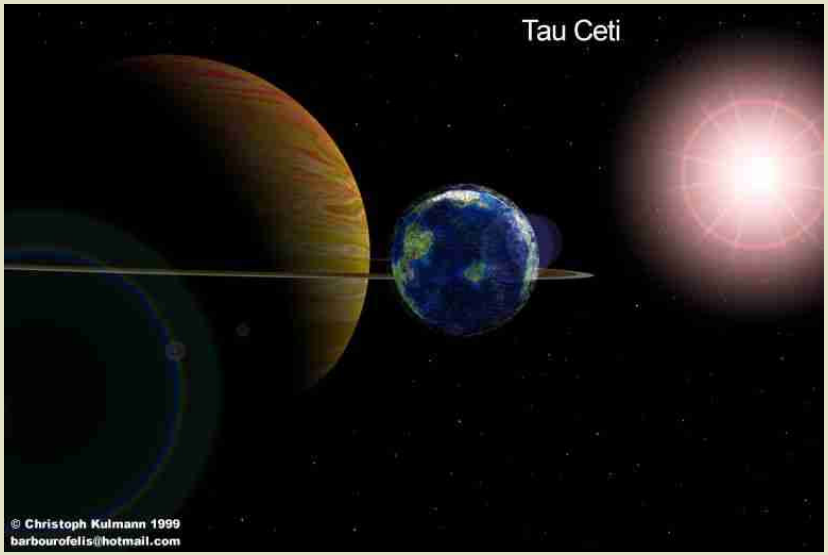
- Moore's Law will have flatlined
- Nano-scale atomic level devices
  - Assuming we solve lithography problem
- Local clock rates ~100 GHz
  - Fastest today is > 700 GHz
- Local actions strongly preferential to global actions
- Non-conventional technologies may be employed
  - Optical
  - Quantum dots
  - Rapid Single Flux Quantum (RSFQ) gates



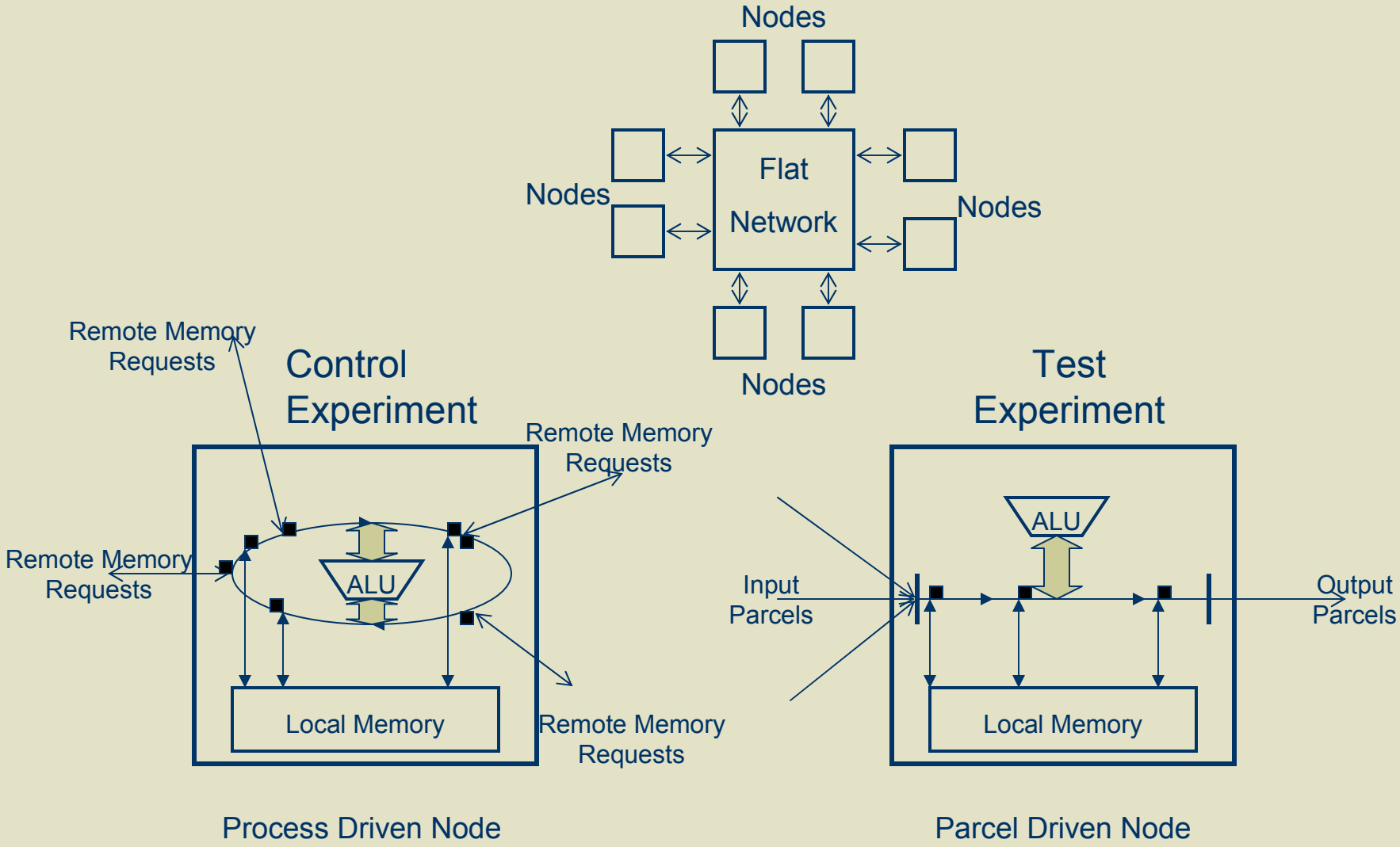


What we will need

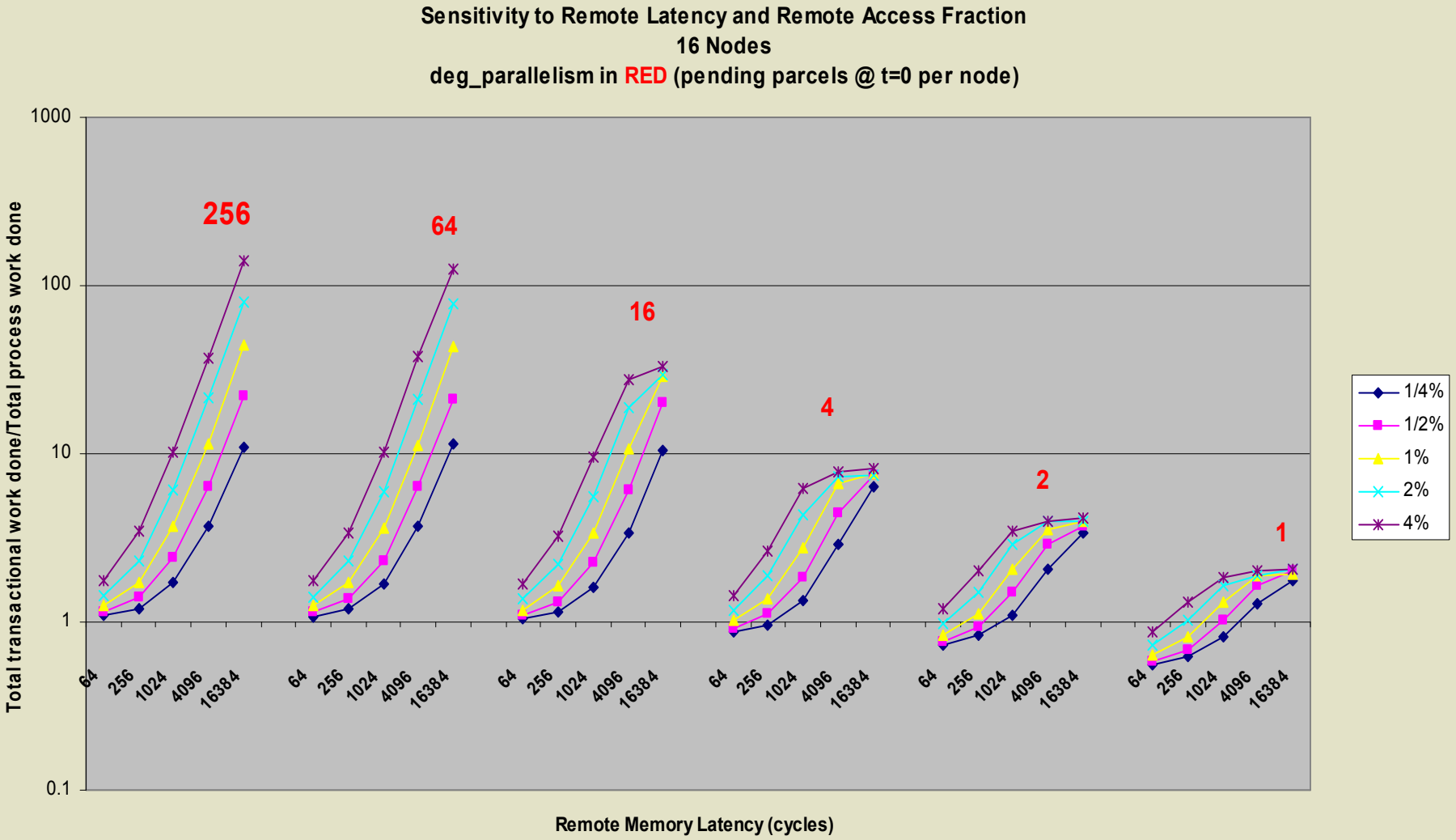
- 1 nano-watt per Megaflops
  - Energy received from Tau Ceti (per m^2)
- Approximately 1 square meter for 1 Zetaflops ALUs
  - 10 billion execution sites
- > 10 billion-way parallelism
- Including memory and communications: 2000 m^2
- 3-D packaging $(4m)^3$
- Global latency of ~ 10,000 cycles
- Including average latency, => 1 trillion-way parallelism



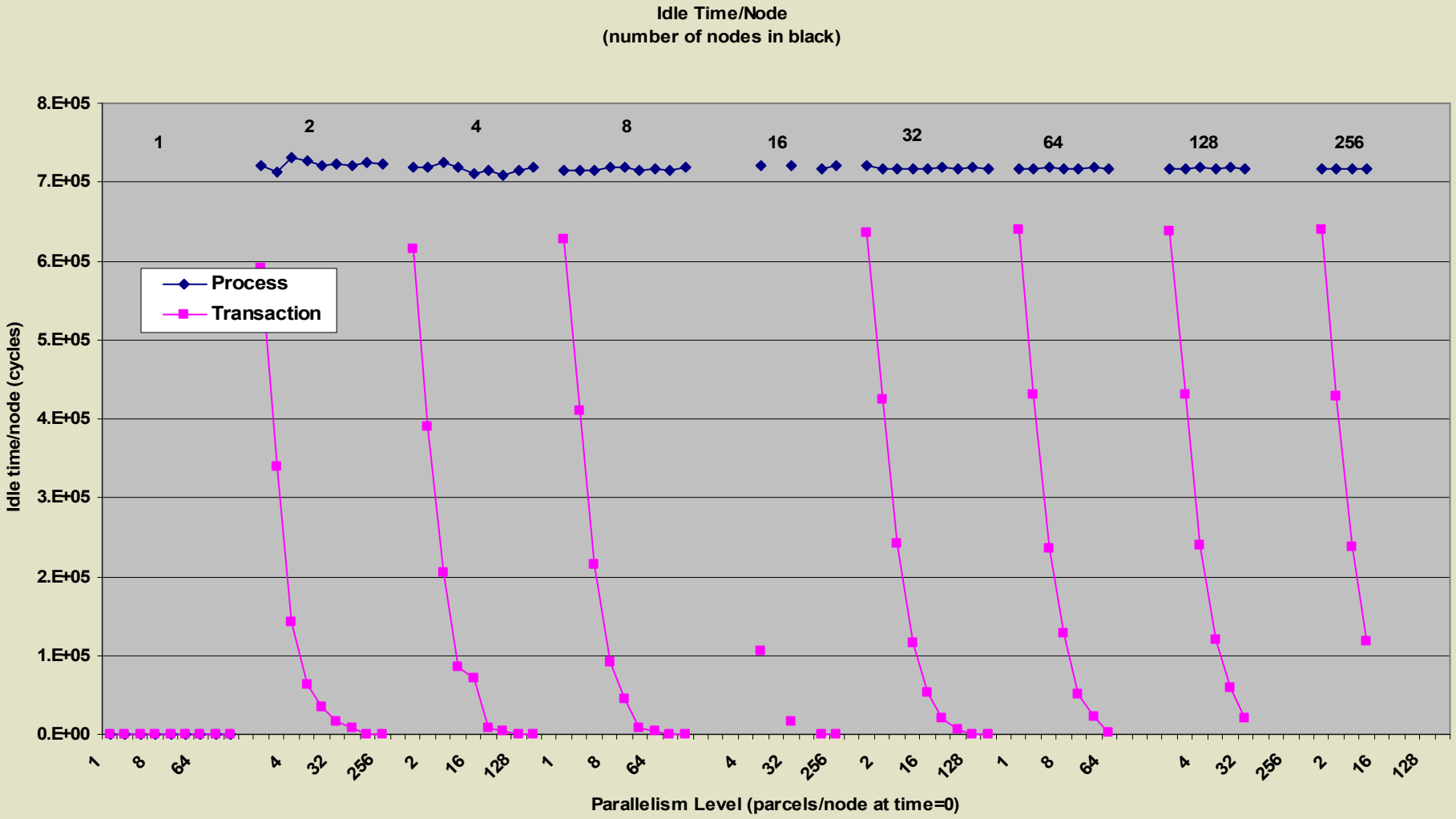
Parcel Simulation Latency Hiding Experiment



Latency Hiding with Parcels
with respect to System Diameter in cycles



Latency Hiding with Parcels
Idle Time with respect to Degree of Parallelism

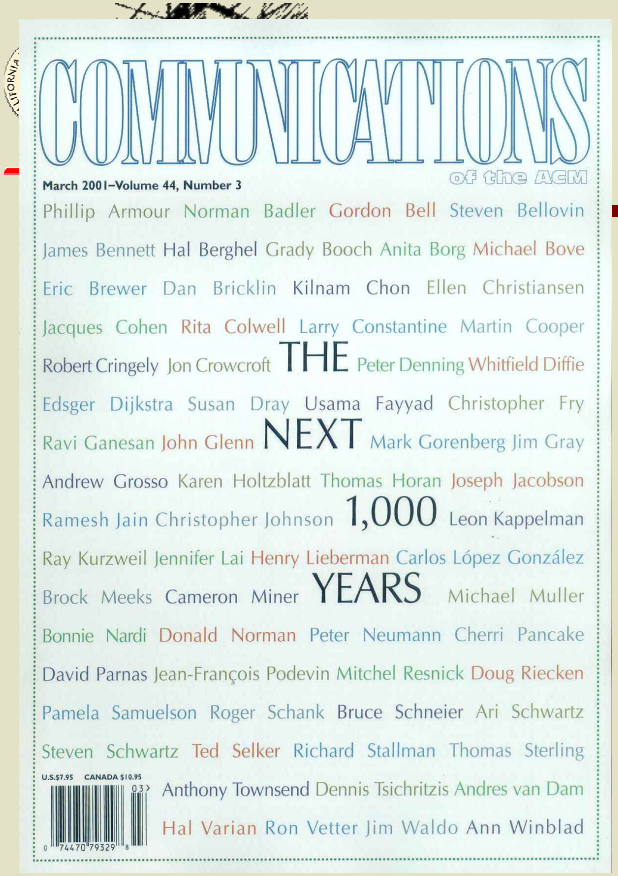


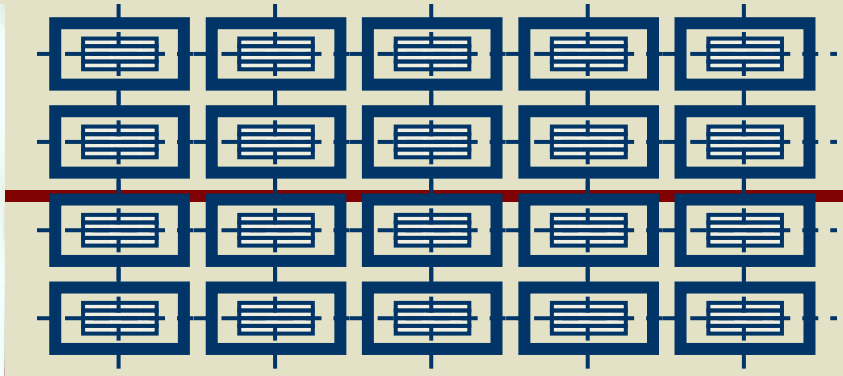
Architecture Innovation

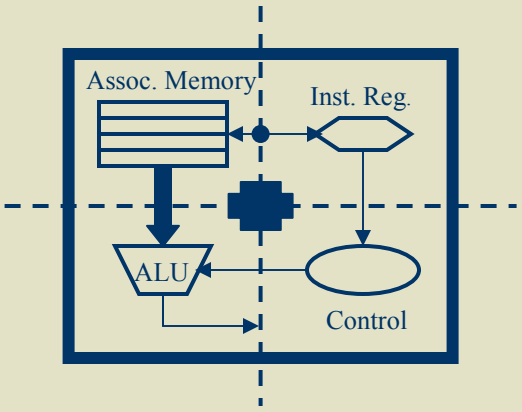
- Extreme memory bandwidth
- Active latency hiding
- Extreme parallelism
- Message-driven split-transaction computations (parcels)
- PIM
  - e.g. Kogge, Draper, Sterling, …
  - Very high memory bandwidth
  - Lower memory latency (on chip)
  - Higher execution parallelism (banks and row-wide)
- Streaming
  - Dally, Keckler, …
  - Very high functional parallelism
  - Low latency (between functional units)
  - Higher execution parallelism (high ALU density)

Continuum Computer Architecture

- Merges state, logic, and communication in single building block
- Parcel driven computation
  - Fine grain split transaction computing
  - Move data through vectors of instructions in store
  - Move instruction stream through vector of data
  - Gather-scatter an intrinsic
  - Very efficient *Futures* for produces-multi-consumer computing
- Combines strengths of PIM and Streaming
  - All register architecture (fully associative)
  - Functional units within a cycle of neighbors
  - Extreme parallelism
  - Intrinsic latency hiding







Continuum Computer Architecture for Exaflops Computation

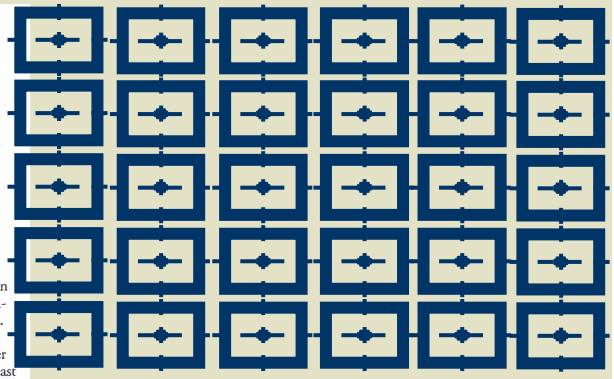
THOMAS STERLING

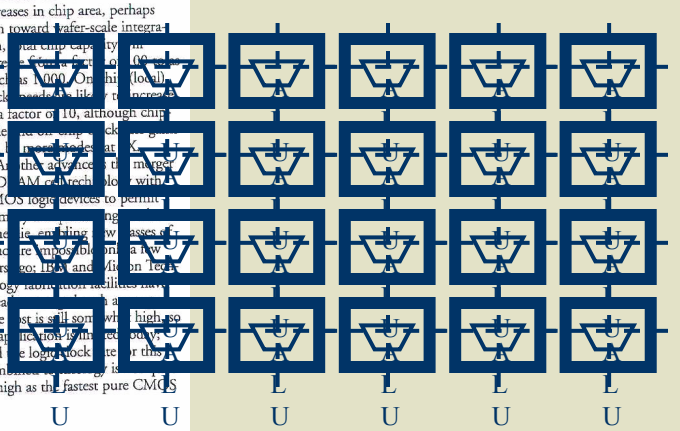
THE ultimate computers in our long-term future will deliver exaflops-scale performance (or greater) and will look very different from today's microprocessors and massively parallel computers. Ironically, however, their alien structures and operational behavior can be inferred from the same technology trends driving development of today's conventional computing systems.

A vision of future computer architectures that are direct extrapolations of current trends is easily inspired by the explosive growth of today's computer performance, price-performance, and applications (driven by Moore's Law for device technology), as well as the more dramatic paradigm shifts brought on by the Internet, the Web, and grids. Yet an examination of these trends also reveals the possibility of something quite different in how we'll organize, design, and fabricate our largest computers in the future. They even set the stage for a revolution in computer architecture that may displace the venerable and highly successful "von Neumann model" and its predominance over the past 50 years.

One class of innovative computing system being explored today by computer scientists at the California Institute of Technology's Center for Advanced Computing Research is the continuum computer architecture (CCA), an ultra-fine-grain uniform structure that approximates a continuous 3D execution medium enabled through next-generation submicron logic devices. Future computers—whether major exaflops engines used to design and simulate everything from controlled fusion reactors to rapid-response medicines, to compact low-power robot brains for autonomous control of spacecraft, airplanes, automobiles, and homes, to embedded smartware in our clothes and bodies—may look less like today's microprocessors and much more like CCAs.

Several concurrent trends in semiconductor and other technologies will force a rethinking of the physical structure and logical operation of parallel computer systems. Lithographic feature size will be driven below 0.05 microns by 2010, increasing the number of devices per unit area by at least an order of magnitude by today's standards. Combined with increases in chip area, perhaps even toward wafer-scale integration, [illegible] by a factor of 10, although chip [illegible] CMOS logic devices to permit [illegible] as high as the fastest pure CMOS





Conclusions

- Zettaflops at nano-scale technology is possible
  - Size requirements tolerable
    - But packaging is a challenge;
  - Latency challenge does not sink the idea
- Major obstacles
  - Power
  - Latency
  - Parallelism
  - Reliability
  - Programming
- Architecture can address many of these
- Continuum Computing Architecture
  - Combines advantages of PIM and streaming
  - Strong candidate for future Zetaflops computer