



Power, Cooling, and Energy Consumption for the Petascale and Beyond

John Shalf: *Lawrence Berkeley National Laboratory*

Bill Tschudi: *Lawrence Berkeley National Laboratory*

Stephen Elbert: *Pacific Northwest National Laboratory*

Rob Pennington: *National Center for Supercomputing Applications*

Andres Marques: *Pacific Northwest National Laboratory*

Tim McCann: *Silicon Graphics Inc.*

Tahir Cader: *ISR Spray Cooling*

Looming Power Crisis



- New Constraints
 - Power limits clock rates
 - Cannot squeeze more performance from ILP (*complex cores*) either!
- But Moore's Law continues!
 - What to do with all of those transistors if everything else is flat-lining?
 - Now, #cores per chip doubles every 18 months *instead* of clock frequency!
- **The "Free Lunch" is over!**

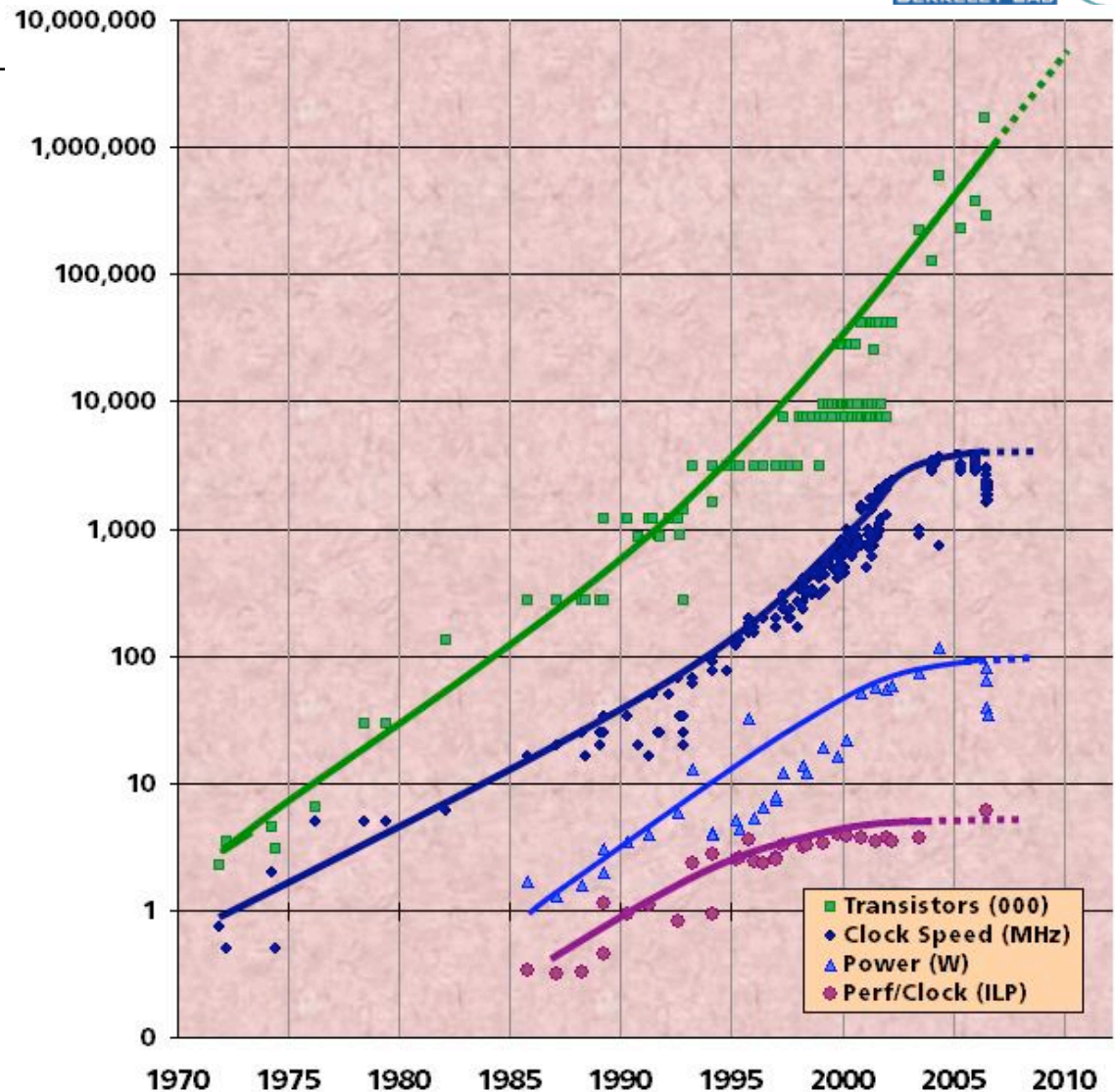


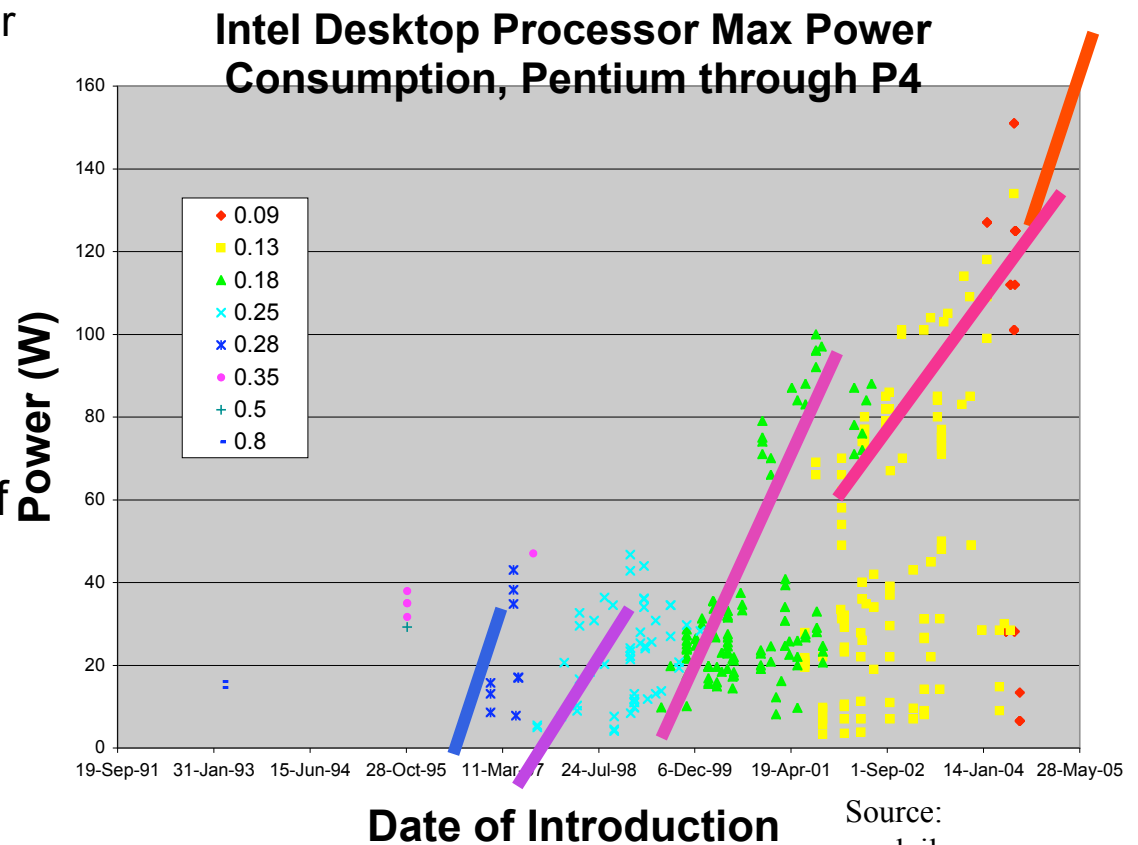
Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith²

Microprocessors: Up Against the Wall(s)



From Joe Gebis

- Microprocessors are hitting a power wall
 - Higher clock rates and greater leakage increasing power consumption
- Reaching the limits of what non-heroic heat solutions can handle
- Newer technology becoming more difficult to produce, removing the previous trend of “free” power improvement



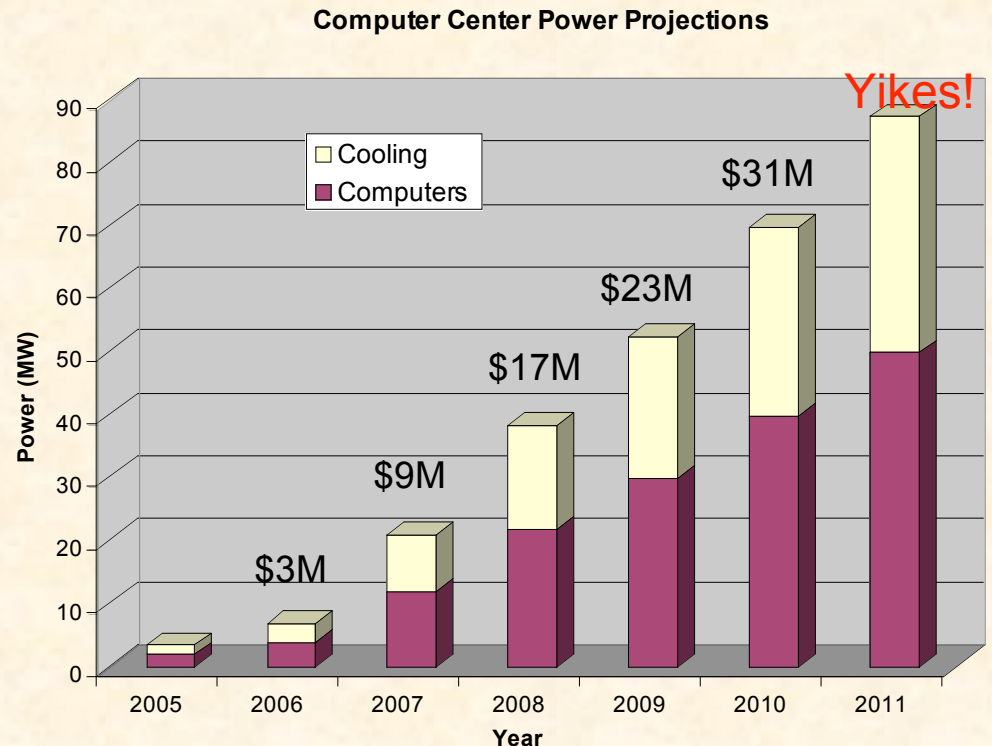
New Design Constraint: *POWER*



-
- Transistors still getting smaller
 - Moore's Law is alive and well
 - But Dennard scaling is dead!
 - No power efficiency improvements with smaller transistors
 - No clock frequency scaling with smaller transistors
 - All “magical improvement of silicon goodness” has ended
 - Traditional methods for extracting more performance are well-mined
 - Cannot expect exotic architectures to save us from the “power wall”
 - Even resources of DARPA can only accelerate existing research prototypes (not “magic” new technology)!

ORNL Computing Power and Cooling 2006 - 2011

- Immediate need to add 8 MW to prepare for 2007 installs of new systems
- NLCF petascale system could require an additional 10 MW by 2008
- Need total of 40-50 MW for projected systems by 2011
- Numbers just for computers: add 75% for cooling
- Cooling will require 12,000 – 15,000 tons of chiller capacity



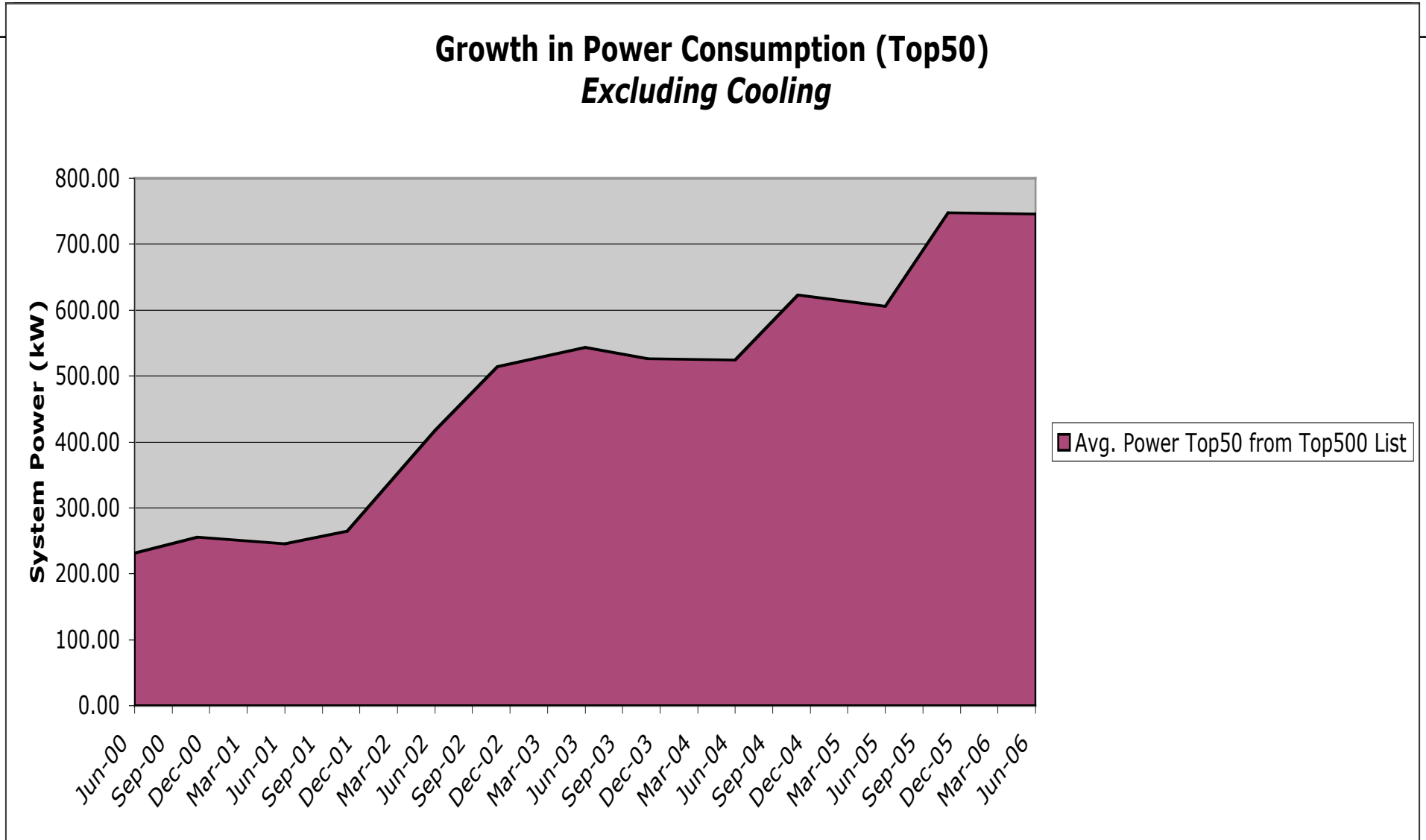
Cost estimates based on \$0.05 kW/hr

Annual Average Electrical Power Rates \$/MWh

Site	FY 2005	FY 2006	FY 2007	FY 2008	FY 2009	FY 2010
LBNL	43.70	50.23	53.43	57.51	58.20	56.40 *
ANL	44.92	53.01				
ORNL	46.34	51.33				
PNNL	49.82	N/A				

Data taken from Energy Management System-4 (EMS4). EMS4 is the DOE corporate system for collecting energy information from the sites. EMS4 is a web-based system that collects energy consumption and cost information for all energy sources used at each DOE site. Information is entered into EMS4 by the site and reviewed at Headquarters for accuracy.

Power Consumption by Top500 Systems



Other Estimates of Power Requirements



- Baltimore Sun Article (Jan 23, 2007): NSA drawing **65-75 MW** in Maryland
 - Crisis: Baltimore Gas & Electric does not have sufficient power for city of Baltimore!
 - expected to increase by **10-15 MW** next year!
- LBNL IJHPCA Study for ~1/5 Exaflop for Climate Science in 2008
 - Extrapolation of Blue Gene and AMD design trends
 - Estimate: **20 MW** for BG and **179 MW** for AMD
- DOE E3 Report
 - Extrapolation of existing design trends to exascale in 2016
 - Estimate: **130 MW**
- DARPA Study
 - More detailed assessment of component technologies
 - Estimate: **20 MW** just for memory alone, **60 MW** aggregate extrapolated from current design trends

The current approach is not sustainable!

Power is an Industry Wide Problem



A screenshot of the CNET News.com website. The header is yellow with the CNET logo and "NEWS.com". Navigation tabs include "Today on CNET", "News", "Reviews", "Compare prices", "How-to", and "Downloads". A search bar is visible. The main article headline is "Power could cost more than servers, Google warns" by Stephen Shankland, published on December 9, 2005. An advertisement for a silver car is shown on the right.

The New York Times "Hiding in Plain Sight, Google Seeks More Power",
by John Markoff, June 14, 2006



**New Google Plant in The Dulles, Oregon,
from NYT, June 14, 2006**

How Big is the Problem?

Numbers represent
U.S. only



- Estimated Computing Power Consumption

- 200 TWh/year

- \$16 billion/year

- Based on .08\$/KWh, closer to \$.10 now (2005)

- Nearly 150 million tons of CO₂ per year

- Roughly equivalent to 30 million cars!

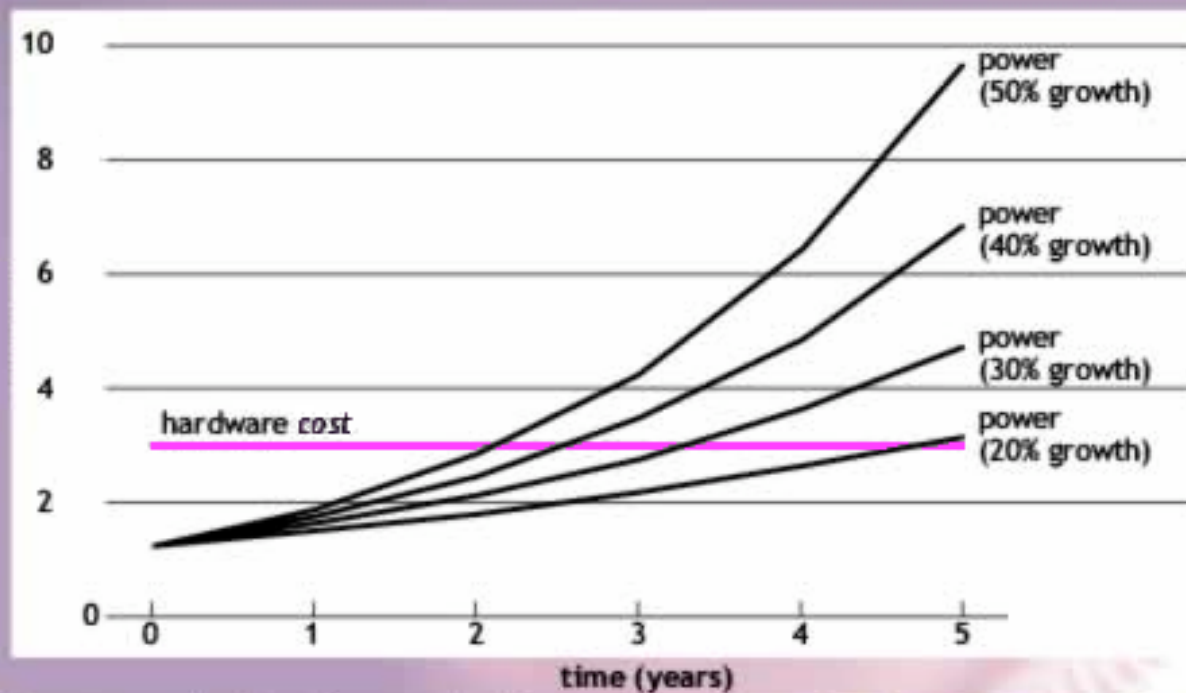
One central baseload power plant
(about 7 TWh/yr)



Cost of Power Will Dominate, and Ultimately Limit Practical Scale of Future Systems



Extrapolation of Hardware and Power Costs for Low-End Servers*



*assumes constant performance/watt over the next five years

FIG 2

Unrestrained IT power consumption could eclipse hardware costs and put great pressure on affordability, data center infrastructure, and the environment.

Source: Luiz André Barroso, (Google) "The Price of Performance," *ACM Queue*, Vol. 2, No. 7, pp. 48-53, September 2005. (Modified with permission.)



Power Efficiency BoF

- Chip Architecture Trends for Power Efficient Computing
- Review Facility Design features for improved power and cooling efficiency
- Discuss cooling technology for future HPC system designs and its impact on facility design
- System architecture features to save power
- Discuss emerging energy efficiency standards and groups
 - ASHRAE
 - Green Grid
 - Green500



More Information

- All of the BoF Talks Online at
 - <http://esdc.pnl.gov>
- More Information on Power Efficient Datacenters:
 - <http://hightech.lbl.gov/datacenters>
- Computer Architecture
 - <http://view.eecs.berkeley.edu/>
 - <http://www.nersc.gov/projects/SDSA/reports>
- Information / Metrics / Standards Bodies
 - <http://www.ashrae.org/>
 - <http://www.thegreengrid.org/>
 - <http://www.green500.org/>
 - <http://www.80plus.org/>
 - <http://www.climatesaverscomputing.org/>



Some Short Remarks on Computer Architecture Trends

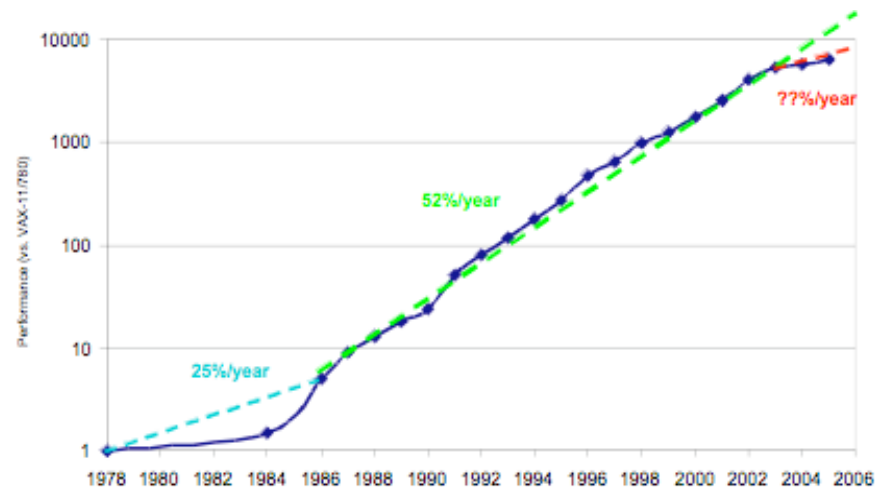
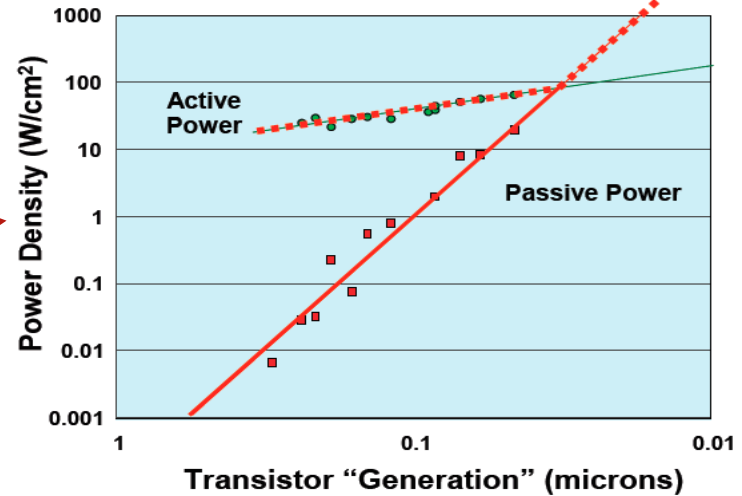
What is Happening Now?



- Moore's Law
 - Silicon lithography will improve by 2x every 18 months
 - Double the number of transistors per chip every 18mo.
- CMOS Power

Total Power = $V^2 * f * C$ (active power) + $V * I_{leakage}$ (passive power)

 - As we reduce feature size Capacitance (C) decreases proportionally to transistor size
 - Enables increase of clock frequency (f) proportionally to Moore's law lithography improvements, with same power use
 - This is called "Fixed Voltage Clock Frequency Scaling" (Borkar '99)
- Since ~90nm
 - $V^2 * f * C \approx V * I_{leakage}$
 - Can no longer take advantage of frequency scaling because passive power ($V * I_{leakage}$) dominates
 - Result is recent clock-frequency stall reflected in Patterson Graph at right



SPEC_Int benchmark performance since 1978 from Patterson & Hennessy Vol 4. 14

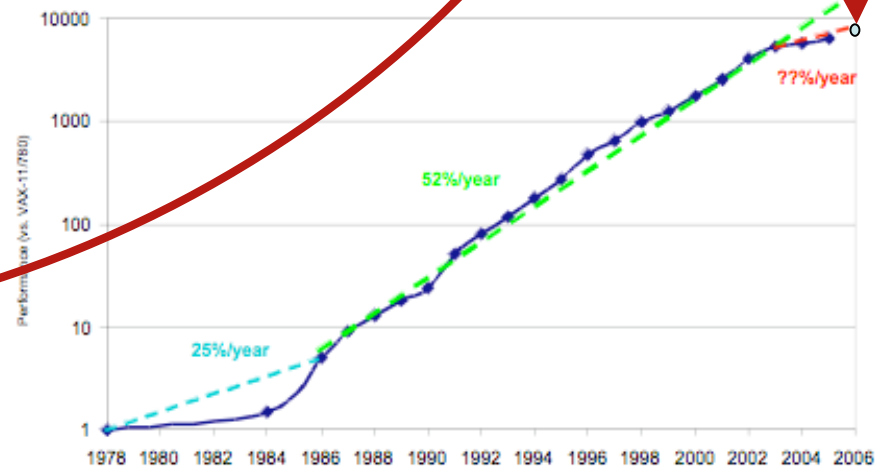
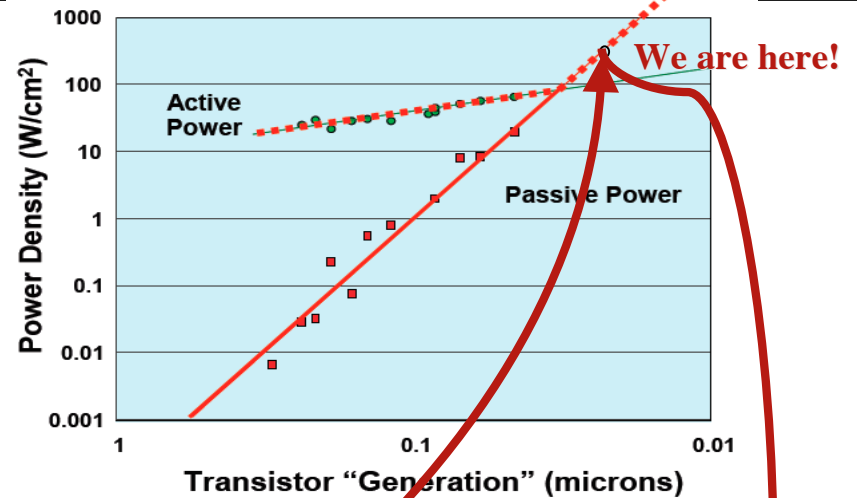
What is Happening Now?



- Moore's Law
 - Silicon lithography will improve by 2x every 18 months
 - Double the number of transistors per chip every 18mo.
- CMOS Power

Total Power = $V^2 * f * C$ + $V * I_{leakage}$
active power passive power

 - As we reduce feature size Capacitance (C) decreases proportionally to transistor size
 - Enables increase of clock frequency (f) proportionally to Moore's law lithography improvements, with same power use
 - This is called "Fixed Voltage Clock Frequency Scaling" (Borkar '99)
- Since ~90nm
 - $V^2 * f * C \approx V * I_{leakage}$
 - Can no longer take advantage of frequency scaling because passive power ($V * I_{leakage}$) dominates
 - Result is recent clock-frequency stall reflected in Patterson Graph at right

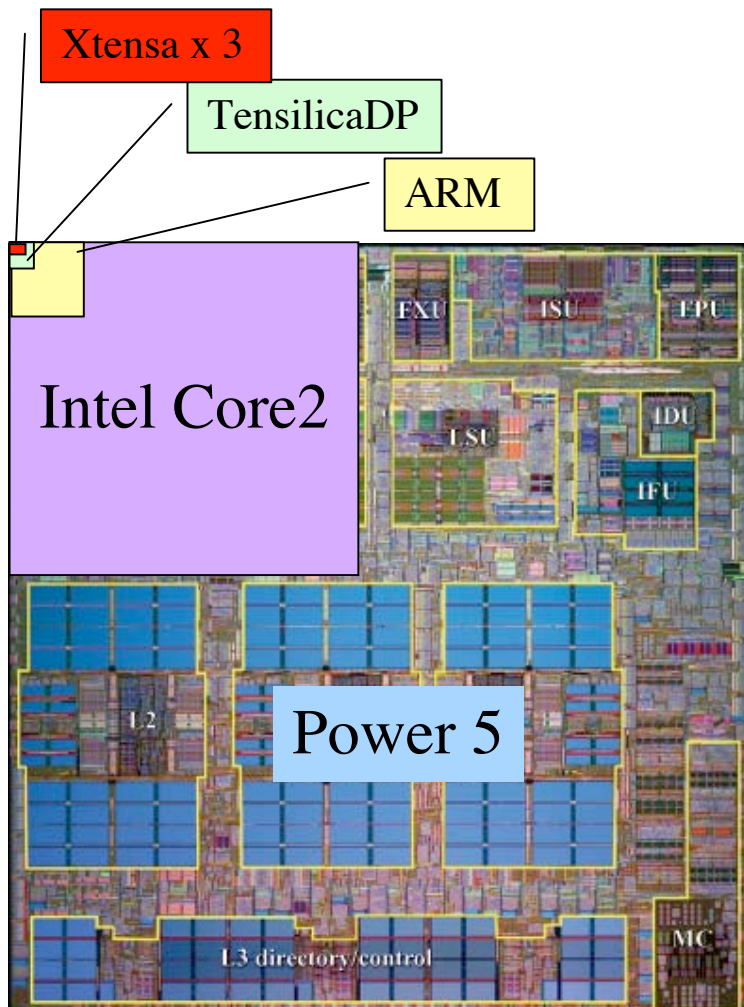


SPEC_Int benchmark performance since 1978 from Patterson & Hennessy Vol 4. 15

Multicore vs. Manycore

- **Multicore: current trajectory**
 - Stay with current fastest core design
 - Replicate every 18 months (2, 4, 8 . . . Etc...)
 - Advantage: Do not alienate serial workload
 - Example: AMD X2 (2 core), Intel Core2 Duo (2 cores), Madison (2 cores), AMD Barcelona (4 cores)
- **Manycore: converging in this direction**
 - Simplify cores (shorter pipelines, lower clock frequencies, in-order processing)
 - Start at 100s of cores and replicate every 18 months
 - Advantage: easier verification, defect tolerance, highest compute/surface-area, best power efficiency
 - Examples: Cell SPE (8 cores), Nvidia G80 (128 cores), Intel Polaris (80 cores), Cisco/Tensilica Metro (188 cores)
- **Convergence: Ultimately toward Manycore**
 - Manycore *if we can figure out how to program it!*
 - Hedge: Heterogenous Multicore

How Small is “Small”



- Power5 (Server)
 - 389mm²
 - 120W@1900MHz
- Intel Core2 sc (laptop)
 - 130mm²
 - 15W@1000MHz
- ARM Cortex A8 (automobiles)
 - 5mm²
 - 0.8W@800MHz
- Tensilica DP (cell phones / printers)
 - 0.8mm²
 - 0.09W@600MHz
- Tensilica Xtensa (Cisco router)
 - 0.32mm² for 3!
 - 0.05W@600MHz

Each core operates at 1/3 to 1/10th efficiency of largest chip, but you can pack 100x more cores onto a chip and consume 1/20 the power

Consider the comparison

From Doug Carmean
Intel Inc.

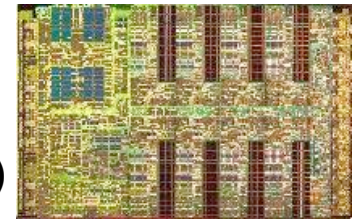
	Traditional Core	Throughput Core	
uArch	Out of Order	In Order	
Size	50	10	mm ²
Power	37.5	6.25	W
Freq	4	4	GHz
Threads	2	4	
Single Thread	1	0.3	Relative Performance
Vector	4 (128-bit)	16 (512-bit)	
Peak Throughput	32	128	GFLOPS
Area Capacity	0.6	13	GFLOPS/mm
Power Capacity	0.9	20	GFLOPS/W



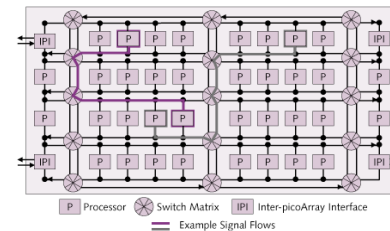
Potential for 20x power efficiency improvement

Convergence of Platforms

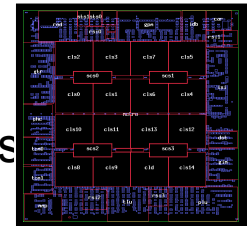
- Multiple parallel general-purpose processors (GPPs)
- Multiple application-specific processors (ASPs)



IBM Cell
1 GPP (2 threads)
8 ASPs

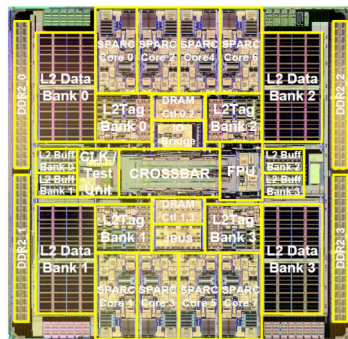
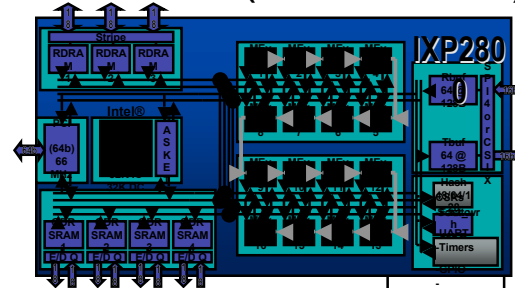


Picochip DSP
1 GPP core
248 ASPs




Cisco CRS-1
188 Tensilica GPPs

Intel Network Processor
1 GPP Core
16 ASPs (128 threads)



Sun Niagara
8 GPP cores (32 threads)

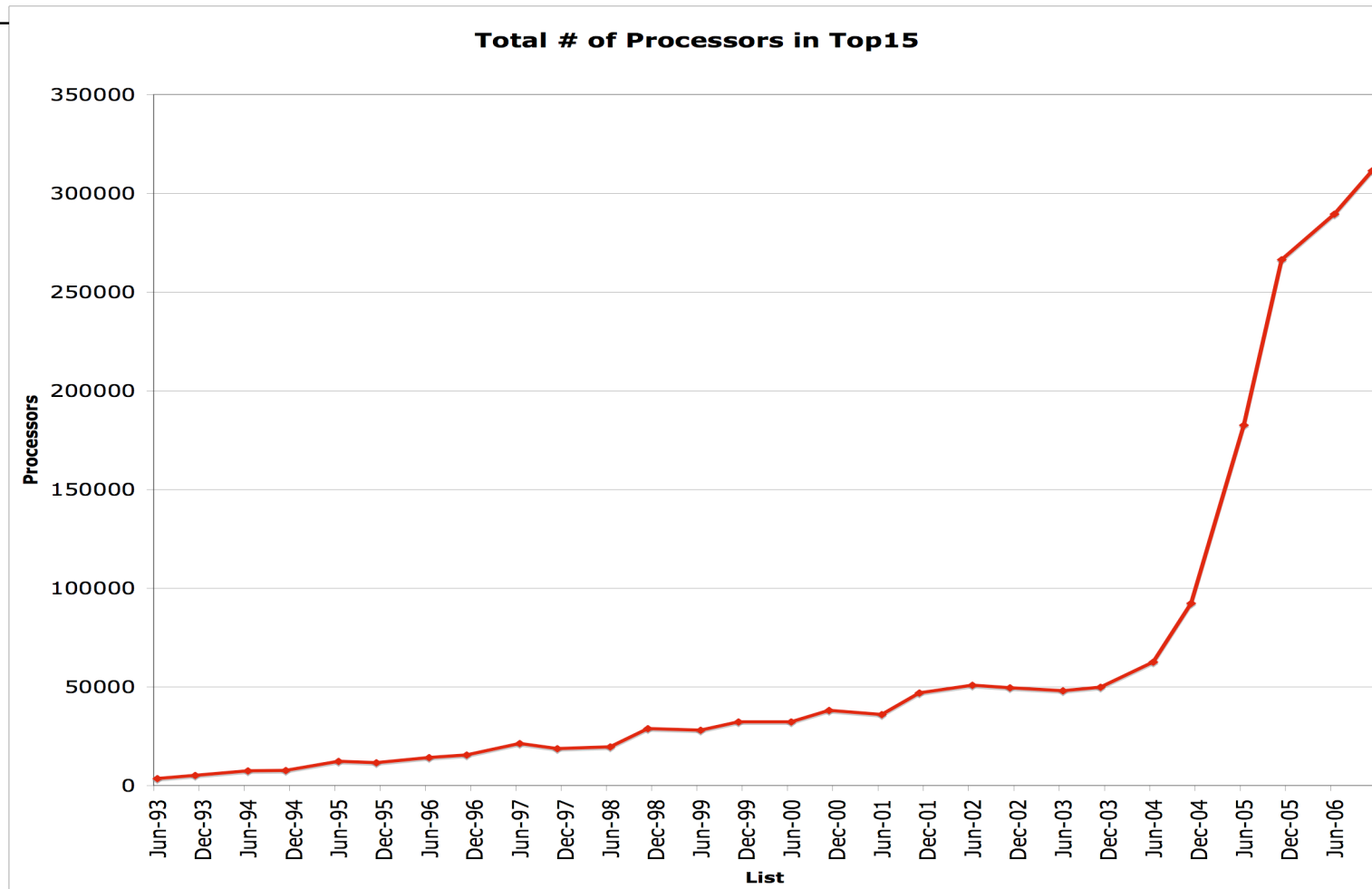


Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm² chip

1000s of
processor
cores per
die

***“The Processor is
the new Transistor”
[Rowen]***

Power Wall Drives Concurrency Increases



**Must ride exponential wave of increasing concurrency for foreseeable future!
You will hit 1M cores sooner than you think!**

Tension between concurrency and power efficiency



- Highly concurrent systems can be more power efficient
 - *Dynamic power is proportional to V^2fC*
 - *Build systems with even higher concurrency?*
- However, many algorithms are unable to exploit massive concurrency yet
 - *If higher concurrency cannot deliver faster time to solution, then power efficiency benefit wasted*
 - *So we should build fewer/faster processors?*

Path to Power Efficiency

Reducing Waste in Computing



- Examine methodology of low-power embedded computing market
 - optimized for low power, low cost, and high computational efficiency

“Years of research in low-power embedded computing have shown only one design technique to reduce power: reduce waste.”

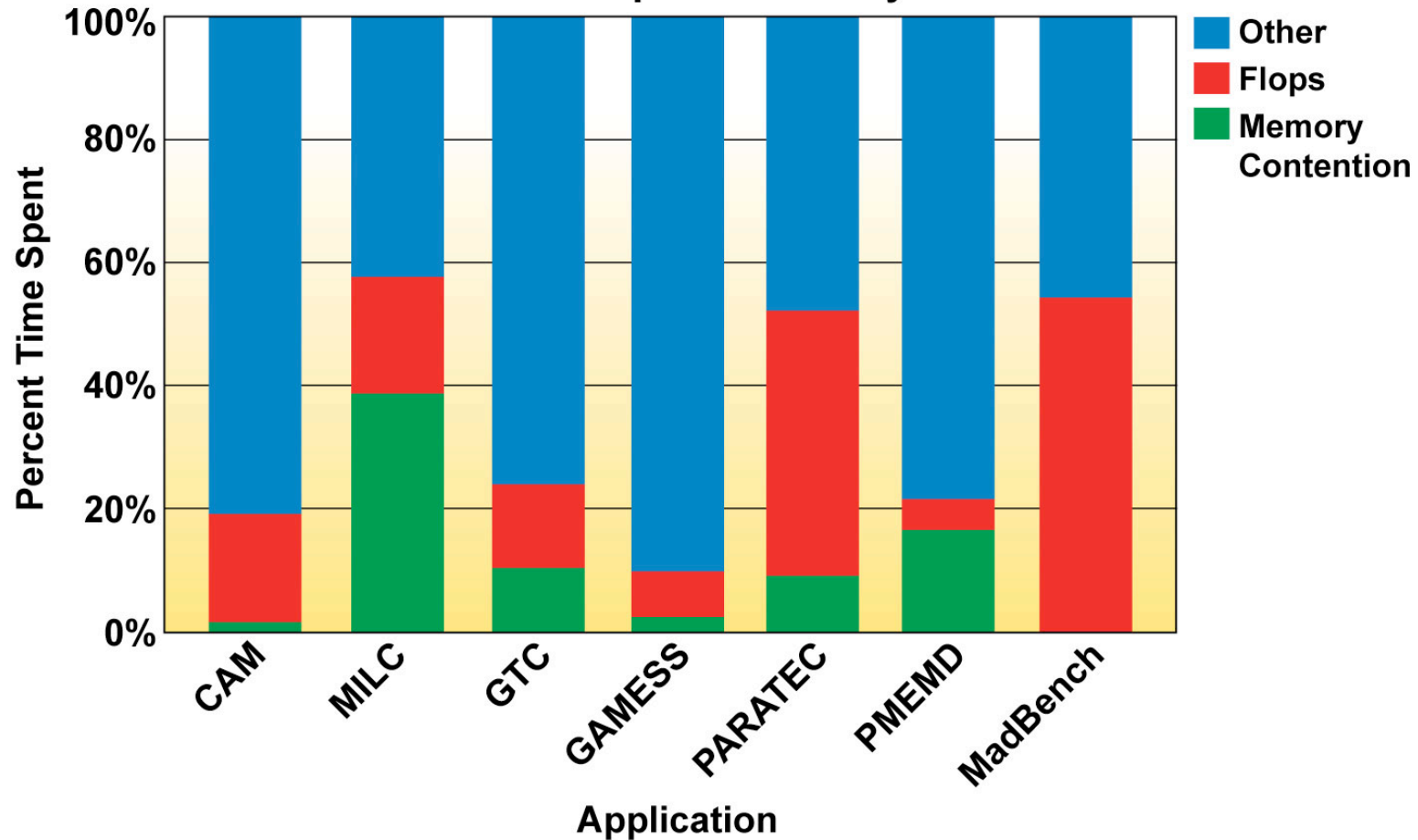
— Mark Horowitz, Stanford University & Rambus Inc.

- Sources of Waste
 - Wasted transistors (surface area)
 - Wasted computation (useless work/speculation/stalls)
 - Wasted bandwidth (data movement)
 - Designing for serial performance

Designing for Efficiency is Application Class Specific

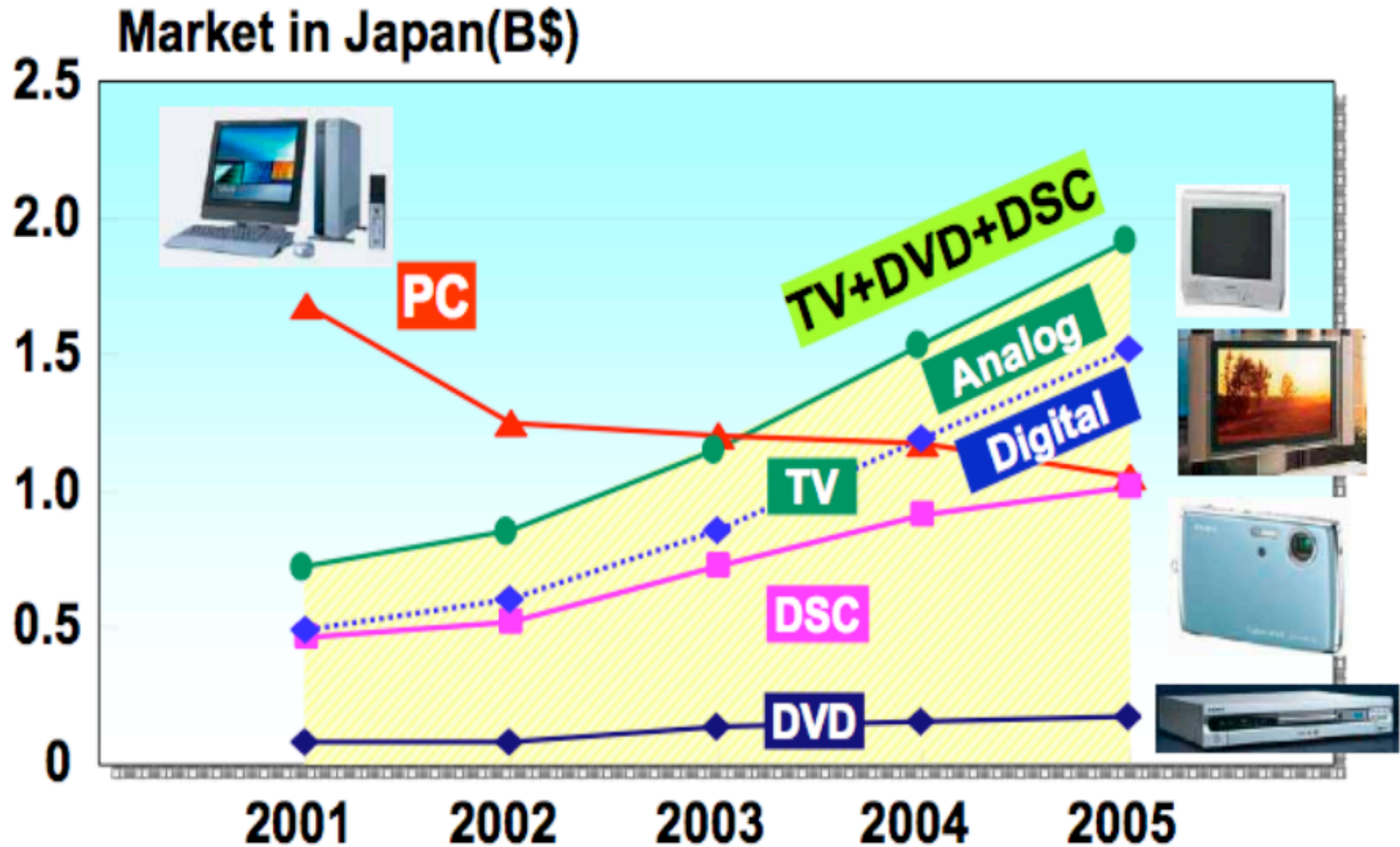


Distribution of Time Spent in Application
In Dual Core Opteron/XT4 System



Consumer Electronics Convergence

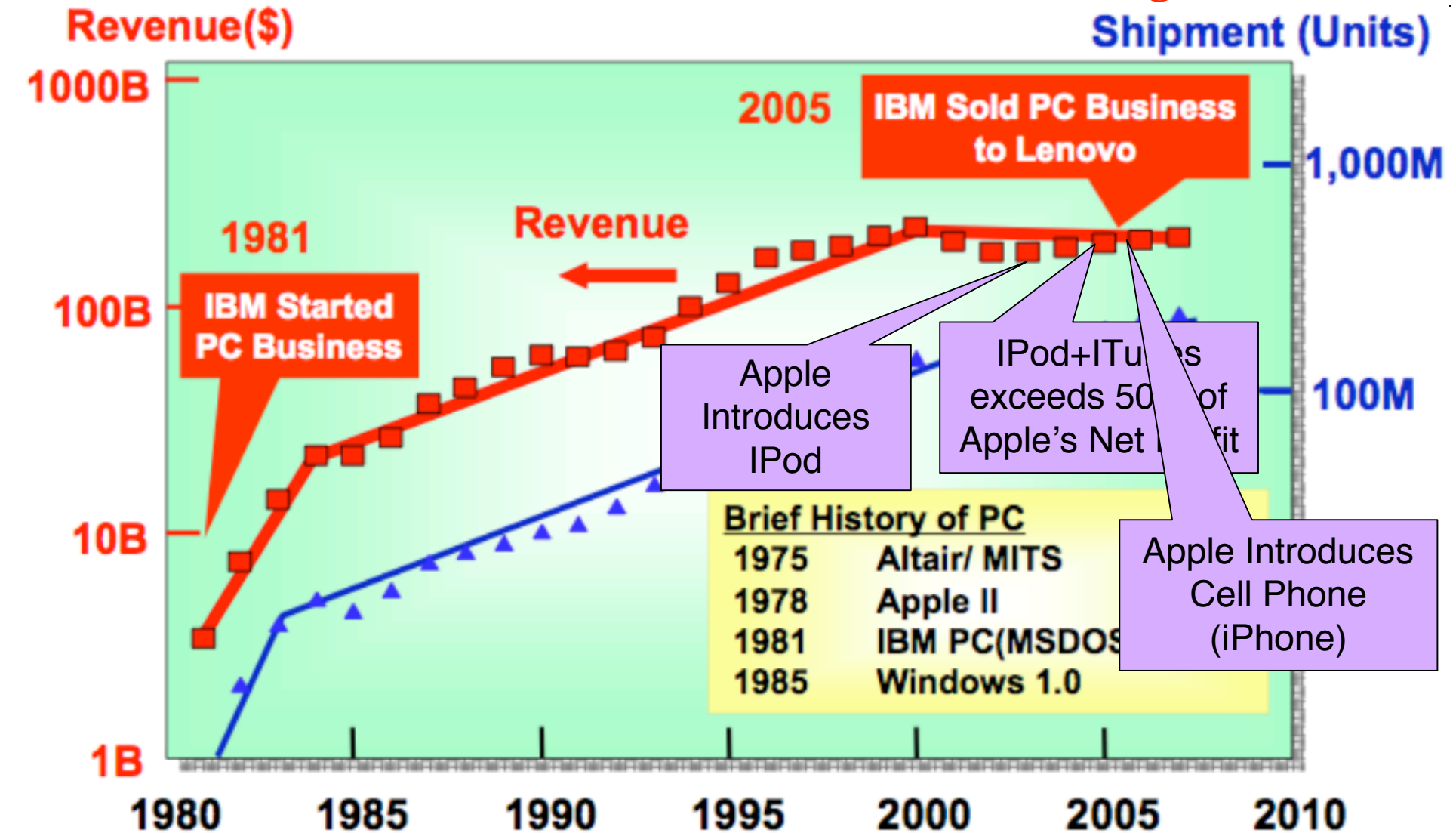
From: Tsugio Makimoto



Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!



From: Tsugio Makimoto

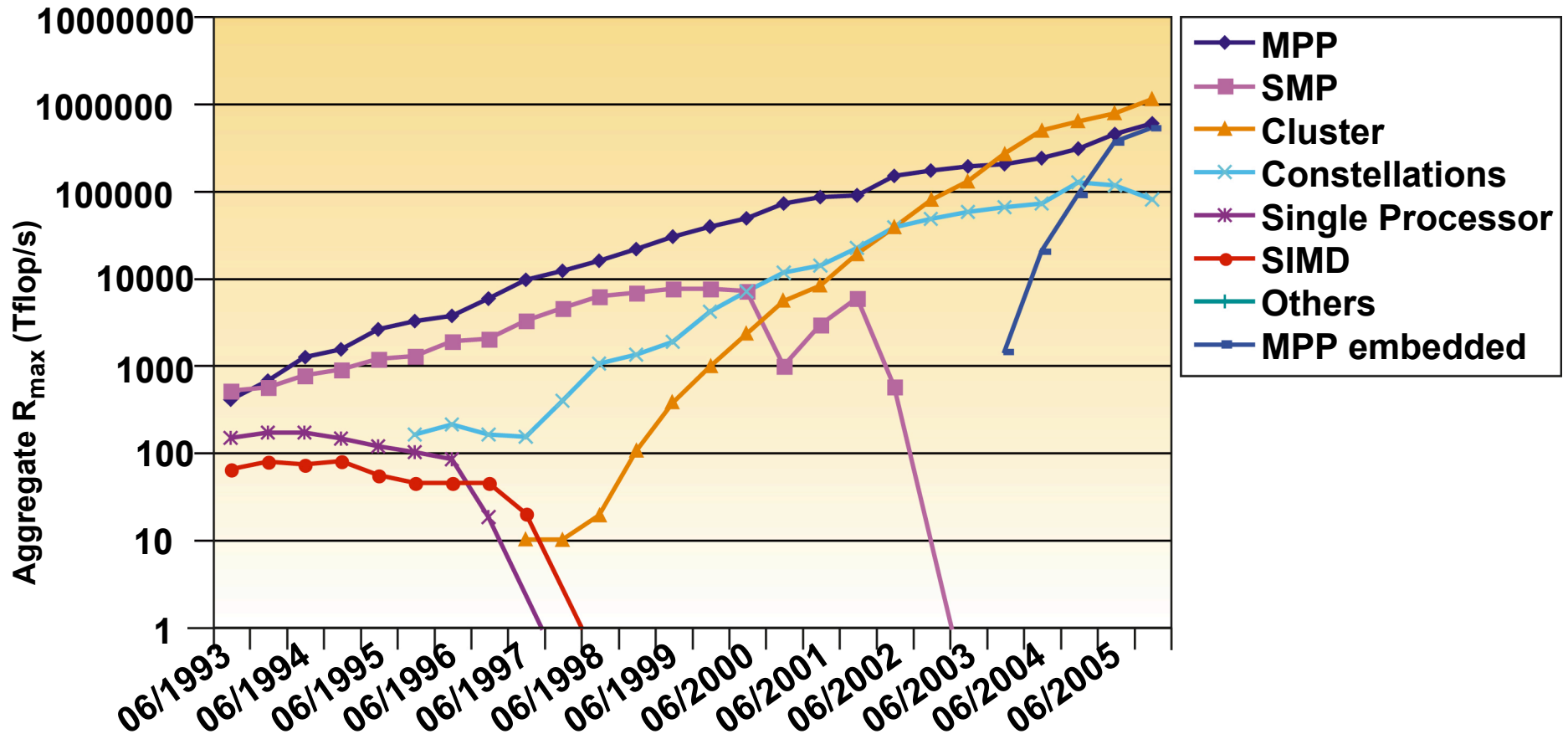


Source: IDC



BG/L—the Rise of the Embedded Processor?

TOP 500 Performance by Architecture



Questions

- Is Multicore really the answer? (sounds boring)
 - FPGAs? Quantum computing?
 - What else might be waiting in the wings
- What about advances in circuit fabrication?
 - SOI, Hafnium doping,
- What about memory?
 - Its starting to consume more memory than CPU cores!
 - Packaging changes (3D Stacking? Optical Interfaces?)



Next Up

Designing Facilities for Power Efficiency