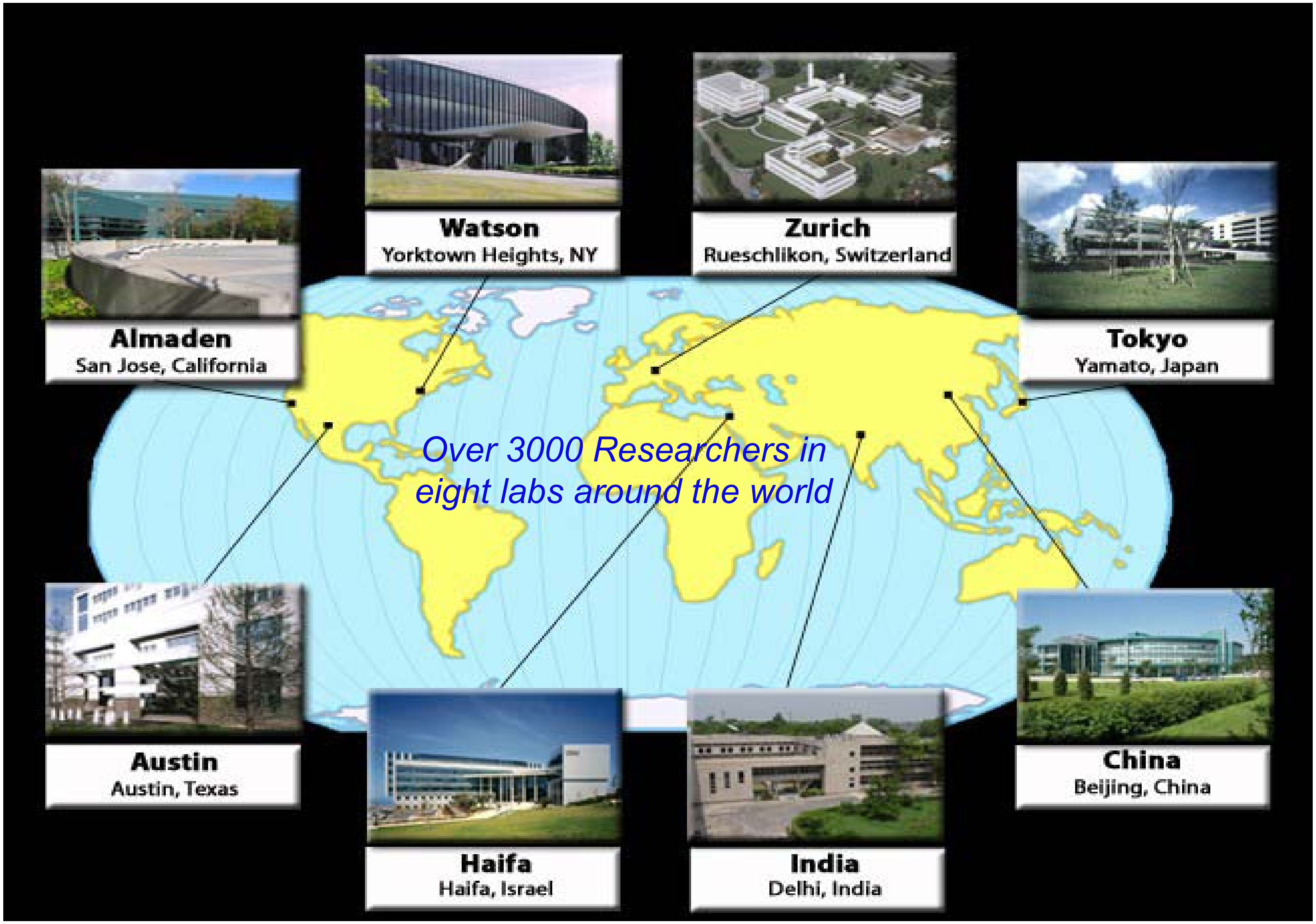
Impact of Technology Trends on Computer Architecture

Jaime H. Moreno
IBM Thomas J. Watson Research Center
Yorktown Heights, NY

August 2008

IBM Research Worldwide

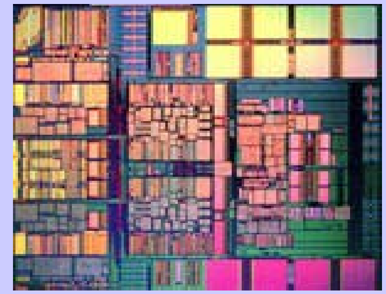


Diversity of Disciplines at IBM Research

Behavioral Sciences

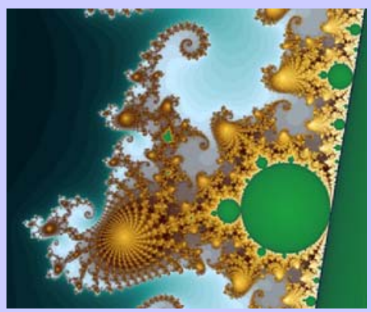
Chemistry

Computer Science

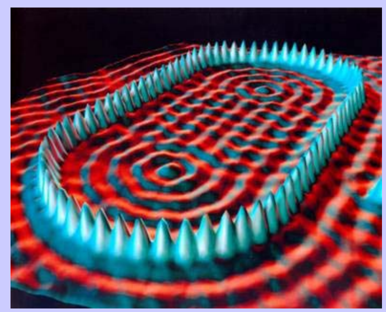


Electrical Engineering

Materials Sciences



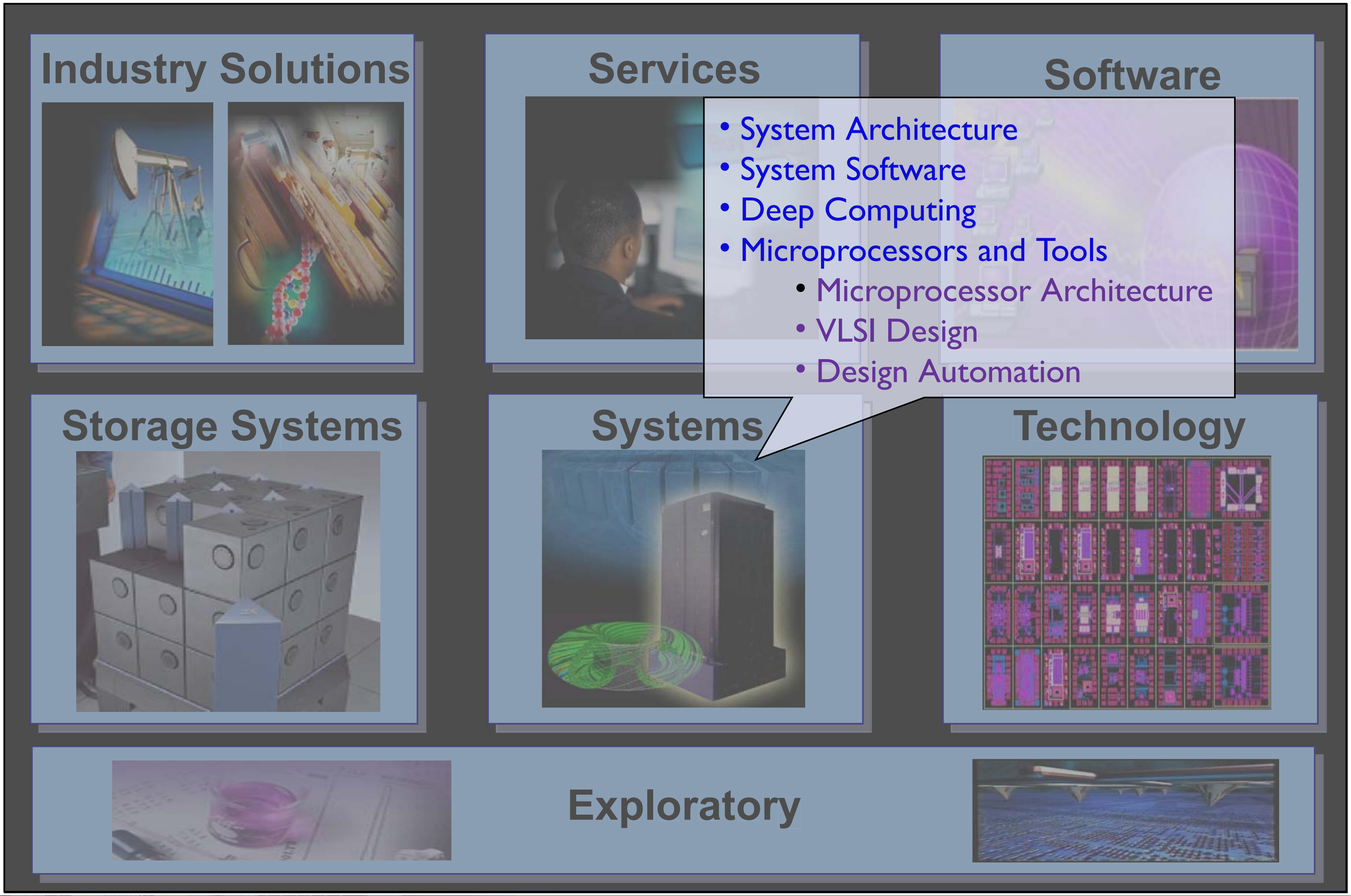
Mathematical Sciences



Physics

Service Science, Management & Engineering

IBM Research's Strategic Thrusts

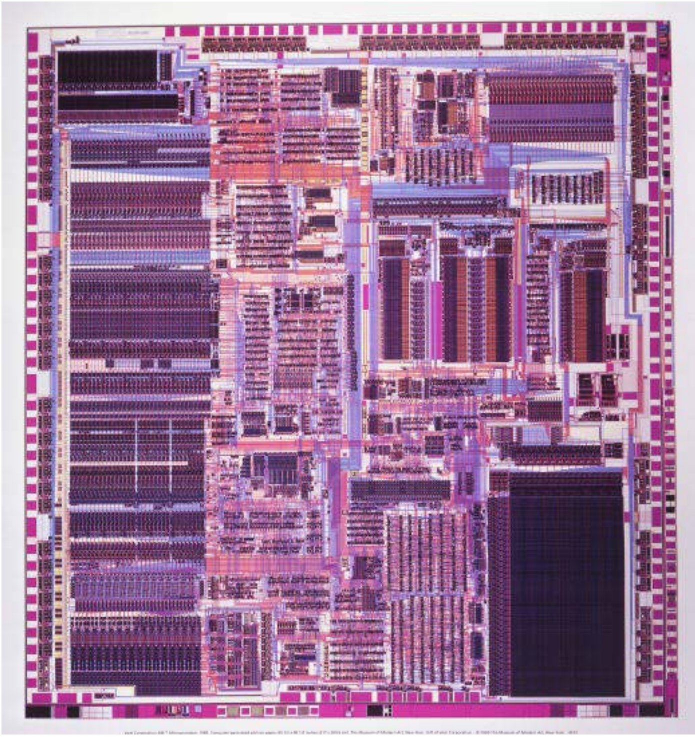


Conversation at a party

- So, what do you do professionally?
- I am a Computer Architect
- (pause ….. silence….)

- Oh, I see… so you design computer cases.
  How come most of the time they are so boring, even ugly?

Architecture vs Chip Architecture

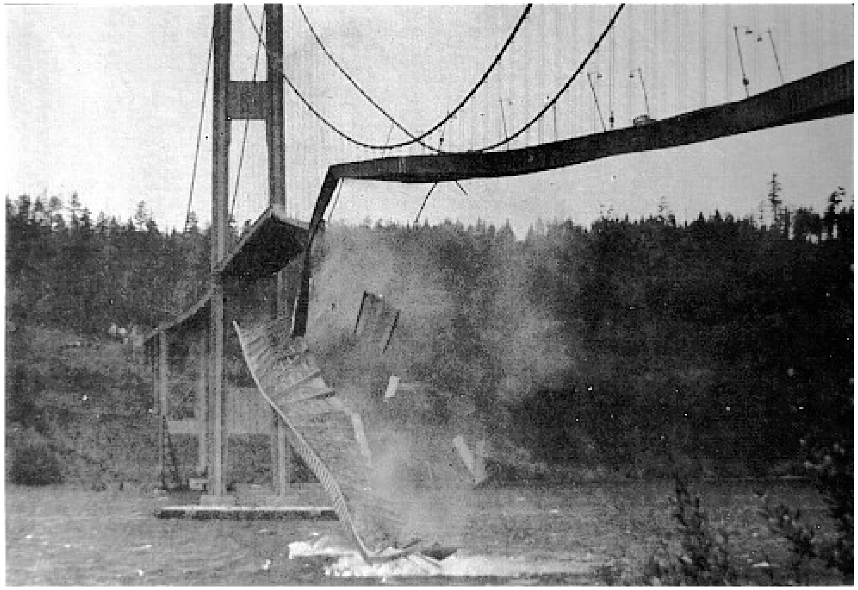


Information Art - Diagramming Microchips
September 6 - October 30, 1990
The Museum of Modern Art, New York

Impact of Technology Trends on Architecture

- Science, engineering and technology are constantly providing new materials, tools, methods, etc., leading to changes in the architecture of buildings
  - Constant evolution of capabilities, enabling all sorts of innovation
- Similarities can also be drawn with other areas
  - Cars, for example

Knowing the Technologies and their Limitations

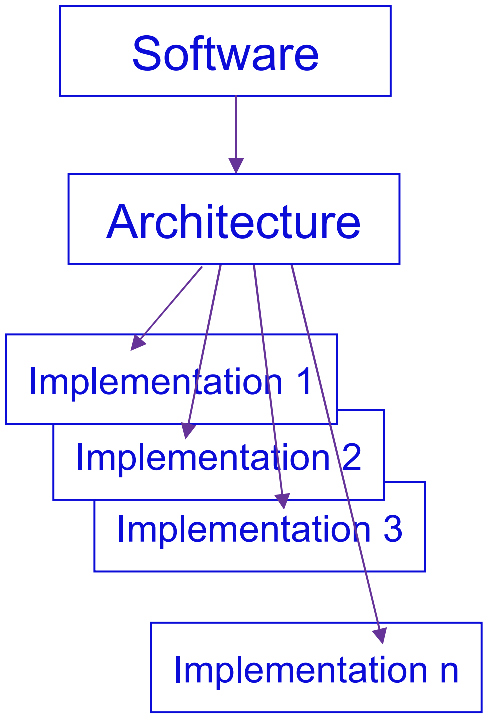


On November 7, 1940, at approximately 11:00 AM, the first Tacoma Narrows suspension bridge collapsed due to wind-induced vibrations. Situated on the Tacoma Narrows in Puget Sound, near the city of Tacoma, Washington, the bridge had only been open for traffic a few months.

"Disasters" may also happen in Computer Architecture, albeit no so dramatically, as a consequence of the use of the technologies

Impact of Technology Trends on Computer Architecture

- What is Computer Architecture?
  - (Valentina already gave us a good perspective)
- In one simple term: a "contract"
  - Conceptually, very similar to the architecture of a building
    - A specification for the "builder" → computer engineers
    - A specification for the "user" → software engineers
- But there are some important differences
  - Technology evolves very fast, so a system looses its advantages in a short period
    - … imagine if that was also the case for homes
  - Although systems change rapidly, expectation is that software migrates from system to system "transparently"
    - Compatibility with "performance" scalability
- Because of the technology changes, a successful "computer architecture" normally has many different "implementations"



Levels of Computer Architecture

- Multiple levels in Computer Architecture
  - Multiple “contracts” throughout a system
  - System, Network, Multiprocessor, Processor, Cache memory, Pipeline, …

- In some cases, we use the term “microarchitecture” to refer to lower levels
  - Microarchitecture closely related to “organization”
  - One “processor architecture” may have multiple “microarchitectures”
  - Each “microarchitecture” may have multiple implementations

- Technology trends impact primarily the implementations, but the effects are often visible at the architecture levels

Technology Trends Example

- A simplistic view
  - Early 70s: Memory was small and slow, compiler technology not mature
    - Complex instruction sets (CISC), "semantic gap," language-specific machines
  - Early 80s: Memory improving rapidly, advances in compiler technology
    - RISC (Reduced Instruction Set) processors emerge
    - Semantic gap replaced by sequences of instructions
    - RISC vs CISC debated extensively
  - Early 90s: Transistor density enables more logic in a chip
    - CISC translated into RISC on-the-fly
    - RISC vs CISC debate became pointless - all RISC internally
  - Early 00s: Transistor density enables even more logic in a chip
    - Processors becoming more CISCy again ..?

- Key observation
  - Changes in technology can trigger major transitions in Computer Architecture
  - Recognizing and anticipating the changes leads to breakthroughs

Examples of Technologies in Computing

- Logic and Main Memory
  - Vacuum tubes
  - Magnetic core memory
  - Semiconductor transistors and memory
  - Integrated circuits: SSI (small), MSI (medium), LSI (large), VLSI (very large)
  - Bipolar, NMOS/PMOS, CMOS
  - Many others
    - Gallium Arsenide, Bubble memories, etc.

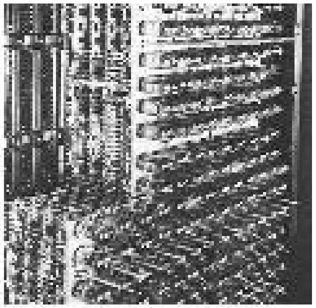
- Storage
  - Magnetic drums, disks, tapes
- Input/output devices
  - Keypads, switches, teletype writers
  - Punched cards, punched tape
  - Dot-matrix printers
  - 32x80 characters display
  - Graphics displays, High-resolution devices
  - etc., etc.

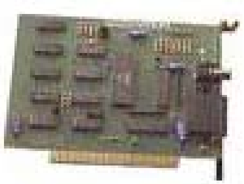
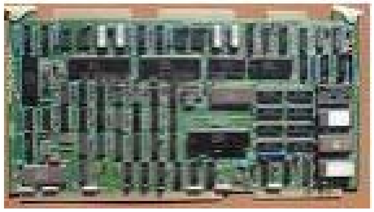
Every major technology innovation has brought major changes to the architecture of computers … and it continues to be so

Technology eras

Examples of Technology Eras in Computing

- Vacuum tube era
  - Very little logic, very little memory
  - Very short time to failure
- Magnetic core memory era
  - Non-volatile
  - Larger capacity, more reliable
- Semiconductor era
  - Much larger capacity (logic and memory)

Technologies define Eras

- Not one but multiple technologies define an era
  - Processor and memory
  - Storage
  - Communications
  - I/O devices
  - ….

Primary focus today (Chip architecture)

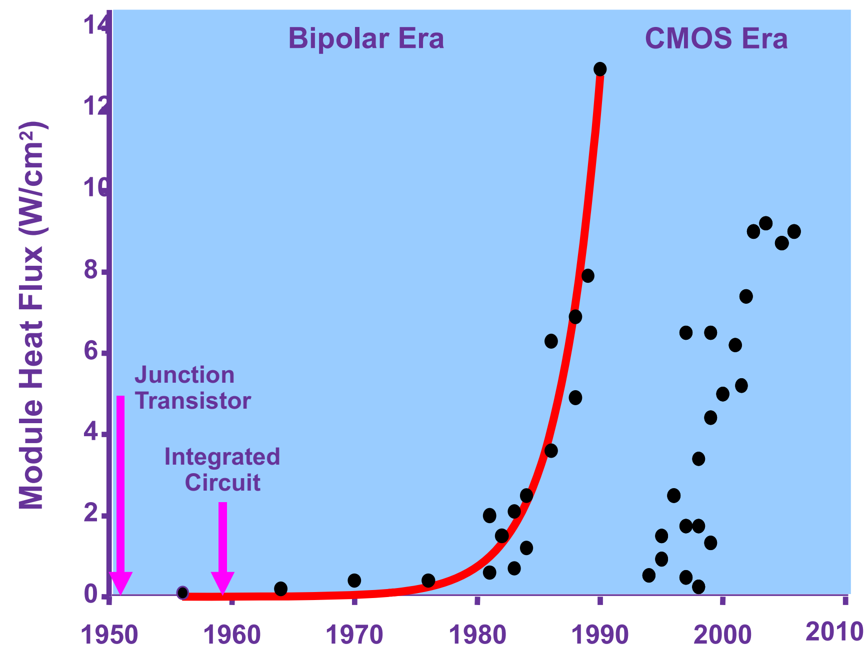
- Consequently, eras have different attributes

- Eras can also be recognized by the usage mode of the computers
  - Batch
  - Interactive
  - Web-driven
  - Real-time driven

Recent Microelectronics Technology Eras

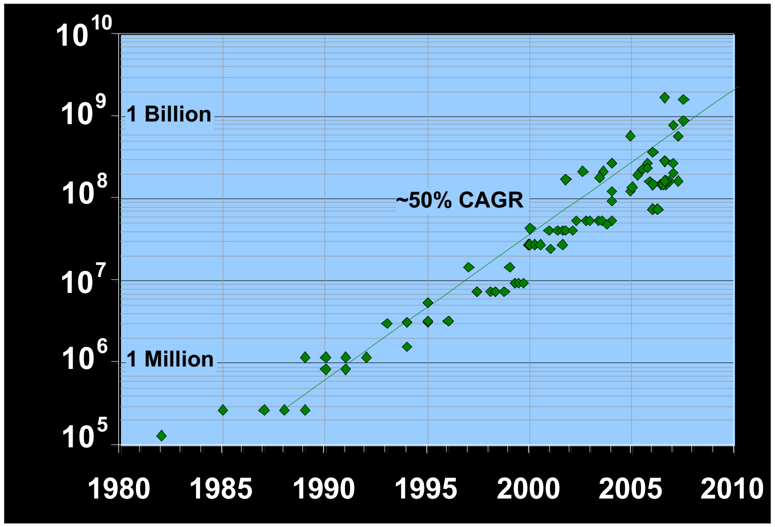
- Bipolar Transistors Era
  - Dominated industry in the early years of modern computers
  - Provided continued improvement in the characteristics of the systems
    - Increased number of transistors in a chip and their speed
  - However, bipolar transistors consume power constantly

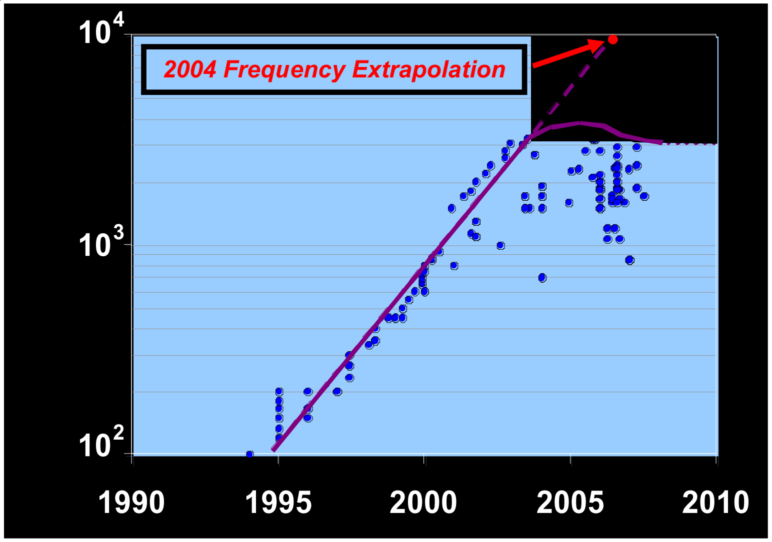
- CMOS Transistors Era
  - Consume power only when there is switching activity
  - Initially, slower than bipolar transistors but eventually became faster
  - Continued trend of improving performance at each generation, without requiring changes to the software
    - Historic trend: 2x every 2 years at roughly constant cost



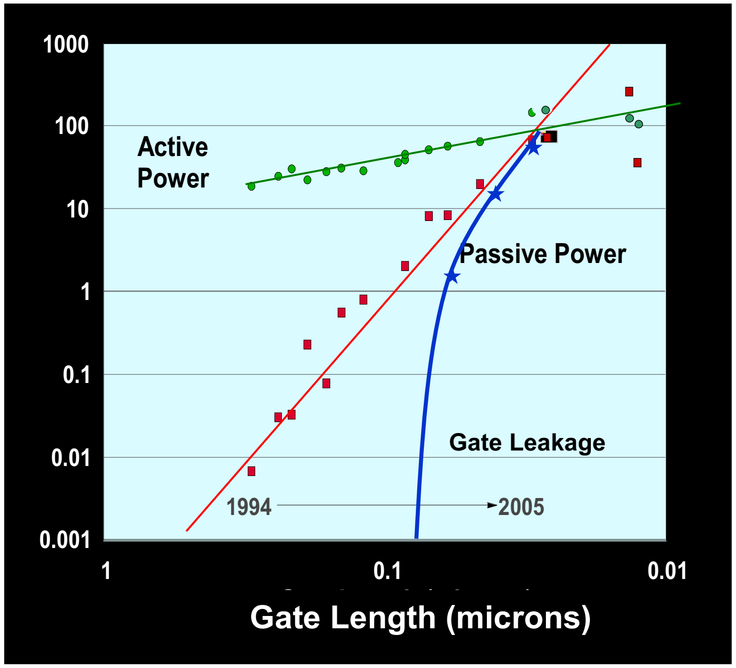
- **Easy migration for software**
- **Widespread adoption of computing technology**

Major Trends: Transistor Density and Frequency



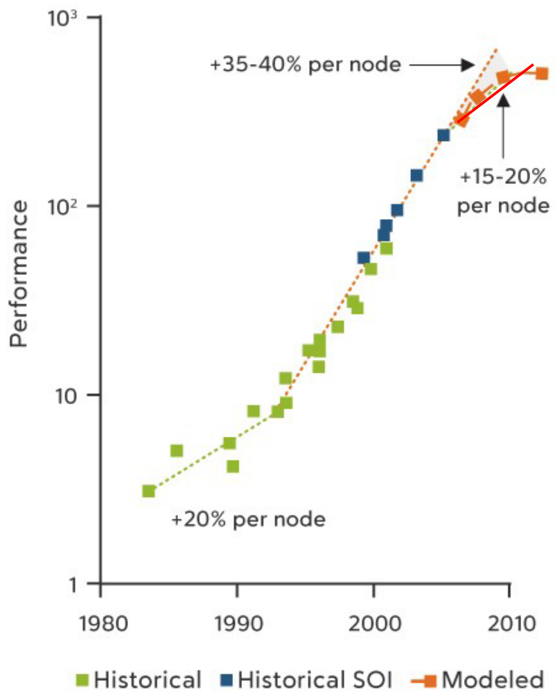


- Where did we go "wrong"?
  - Explosion of leakage current

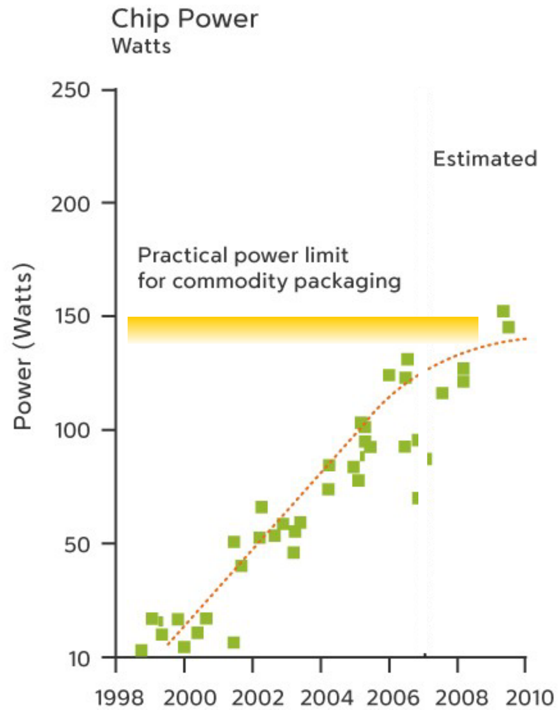


CMOS Era: Benefits from Technology Scaling are Diminishing

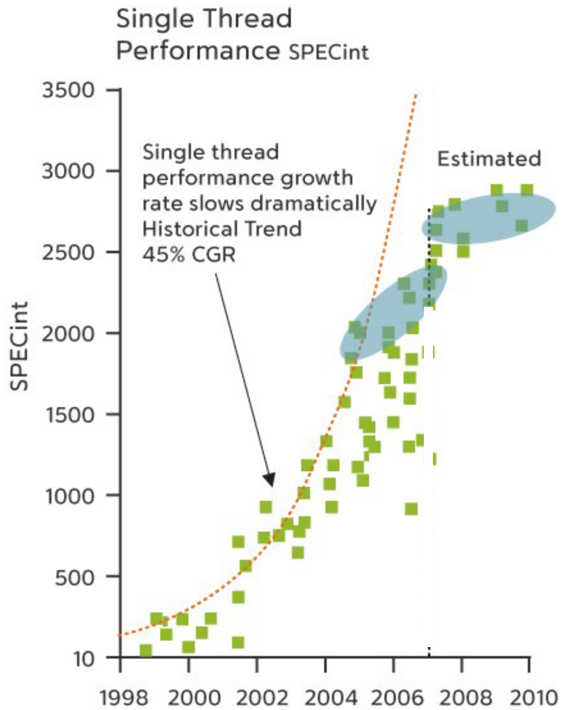
Transistor performance scaling to continue, but at a slower rate



Power is limiting practical performance



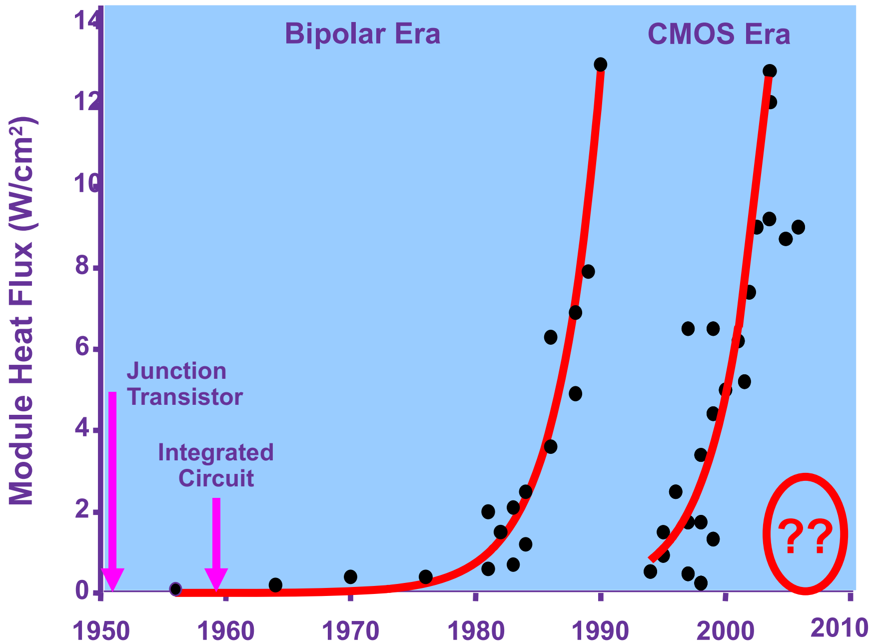
Single thread performance is slowing



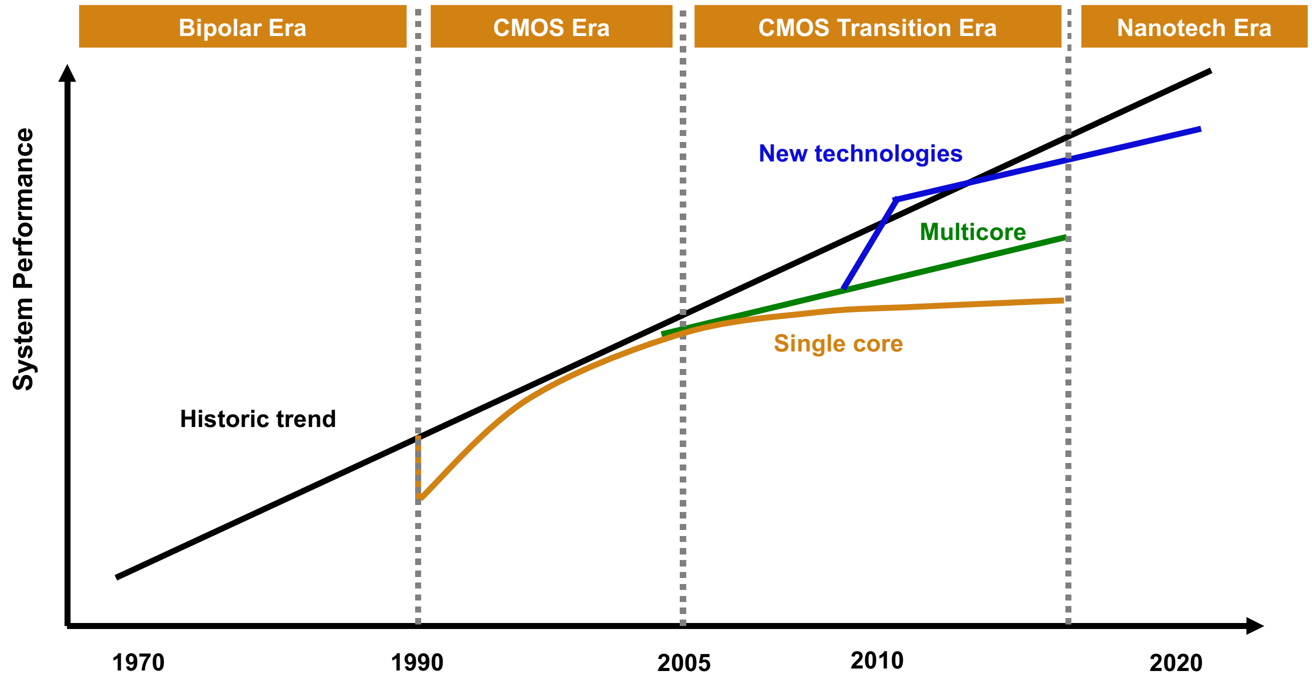
No more "free ride" for software …

Didn't we have this problem in the past…?

- Bipolar Era symptoms are back …
- … but there is no other semiconductor technology ready to be used this time
- … at least not for a number of years
  - Nanotubes
  - Quantum computing
  - Molecular computing
  - ….



A New "Technology" Era - The "CMOS Transition" Era



- CMOS scaling alone can no longer provide simultaneous improvements in all product metrics (performance, cost, power, software compatibility) at the historic rate.
- New technologies, systems and software evolution will be required to achieve performance and cost improvements over the next decade.
- Computer architectures will have to drive the transition

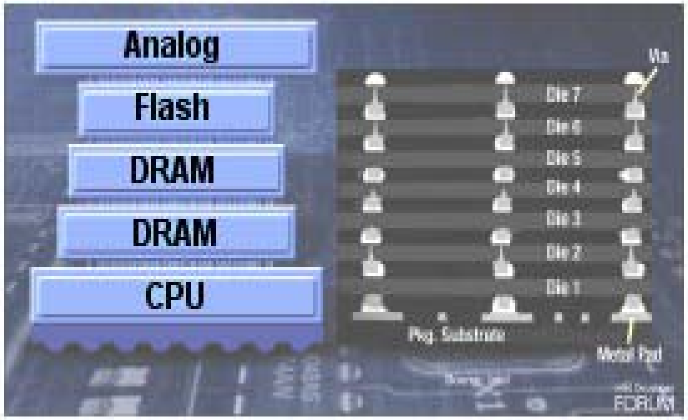
Where are we today?

- CMOS density continues roughly at historic rate while operating frequency is basically flat, because of power limitations
  - We can have more transistors per chip but cannot have them run faster
  - If they run faster, we cannot use all of them at the same time
- Single-thread performance improving slowly
- Transitioning to multicores, accelerators, specialized systems
  - Software changes required
  - Leverage point is moving "up the system stack"

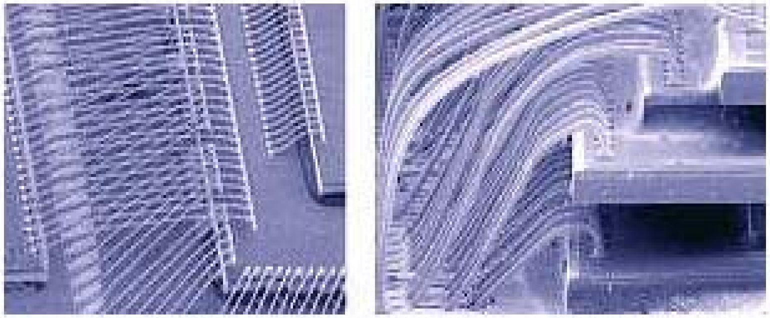
- Technologies other than CMOS needed to compensate the trends
  - Semiconductor materials and structures, lithography, design layout
  - Operation at ultra-low voltage, advanced power management
  - Cooling
  - 3D integration
  - New memory technologies, denser and lower power, non-volatile
  - On-chip optical communications (photonics)
  - Heterogeneous and specialized engines and systems

Why 3D Integration?

- Multi-core, multi-threaded and multi-image (virtualization) chips are stressing memory bandwidth and on-chip memory capacity
- Advances in technology providing much higher I/O densities and bandwidth between chips
- Mix different types of chips in one package
- Design challenges: thermal, package I/O, etc.
- Impact to architecture (example)
  - No longer constrained by 2D organization
  - What if we had Gigabytes of cache memory on chip, and could transfer several Kilobytes per cycle across cache levels?



Solder bumped chip stacked 3D

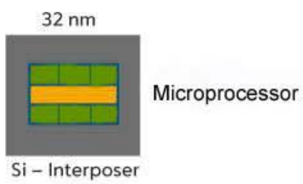


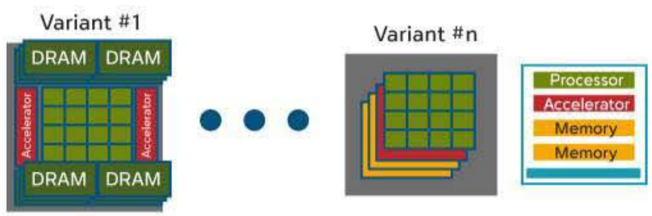
Wire bonded chip stacked 3D

3D Integration

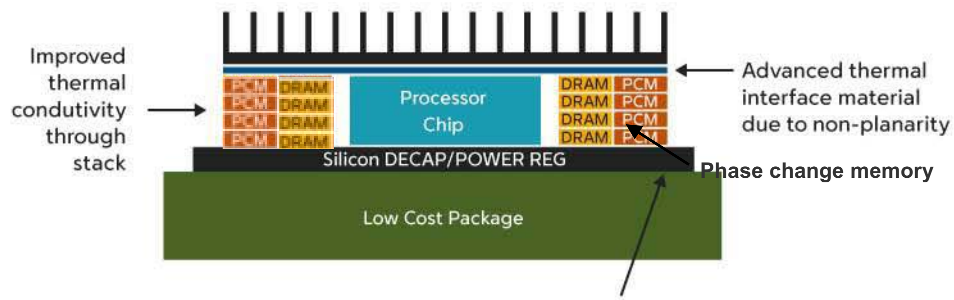
- Restructuring the architecture of the chip / node

3D allows many scenarios of stacking:
- Memory
- Cores
- Accelerators



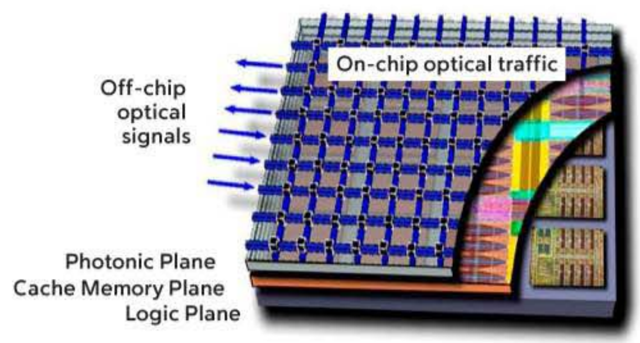


New 3D processor designs



New signaling technologies to reduce total module power to ~100 W
As high as 250 W with current driver technology

Potentially leading to a Node in a Socket



New Memory Technologies

- Phase-change Memory (PCM) attributes:
  - Fast: Tens of nano-seconds
  - Low-Power: Non-volatile storage
  - Dense: Multi-bit, projected to be comparable to disks
- Impact to architecture (example)
  - What if we had Terabytes of main memory, but the number of write operations are limited?

	DRAM	PCM	NAND FLASH	ENTERPRISE DISK
Read access time (us)	0.05	0.1	20	5000
Random Write Access (us)	0.05	0.1	1500	5000
Device Capacity (GB)	0.5	140	32	500 - 2500
Device Bandwidth (MB/s)	1000	1000	20	150
Endurance	10^{15}	10^{9} - 10^{12}	10^{5}	10^{12}
Device Power (W)	0.2	0.1	0.1 - 0.2	10-20

Legend: Best Barrier

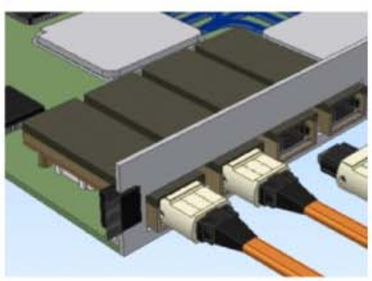
Optical Interconnect Technology

- Impact to architecture (example)
  - What if we could transfer data within a system 100 times faster?

2008

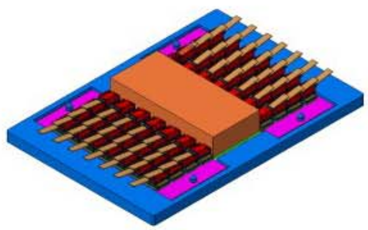
5G Active Cable

5G optics at rack edge

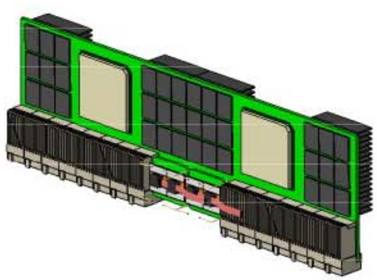


2010

10G Optics on chip



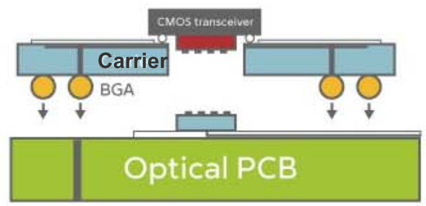
10G Optics mid-card



2014

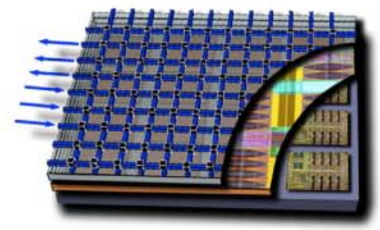
10G–20G Optics on card

Possible technologies:
Fiber, waveguide flex
or WG on card



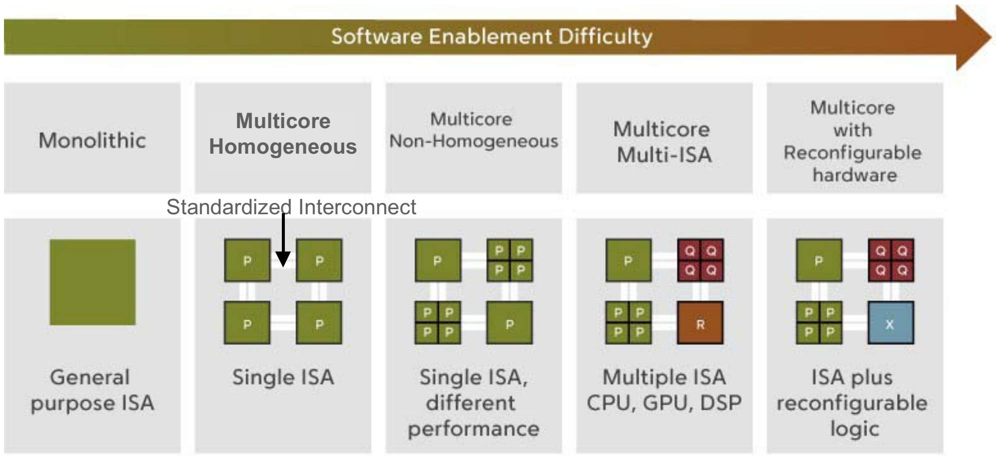
2018

10G–20G+ Optics w/3D
Integrated Photonics



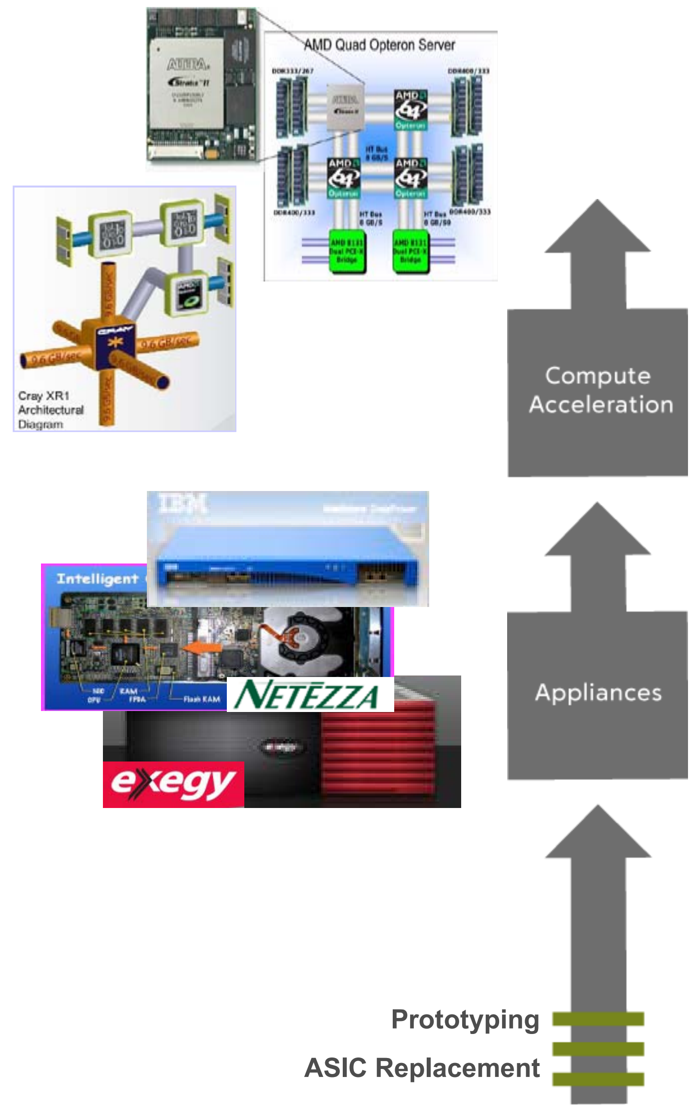
Heterogeneous Systems

- Emergence of heterogeneous components in processor and systems
  - Recently announced "Roadrunner" supercomputer has Cell and x86 chips
    - Fastest and most power efficient computer in the world
- Impact to architecture (example)
  - What is the architecture of a system with increased heterogeneity (hardware and software)?



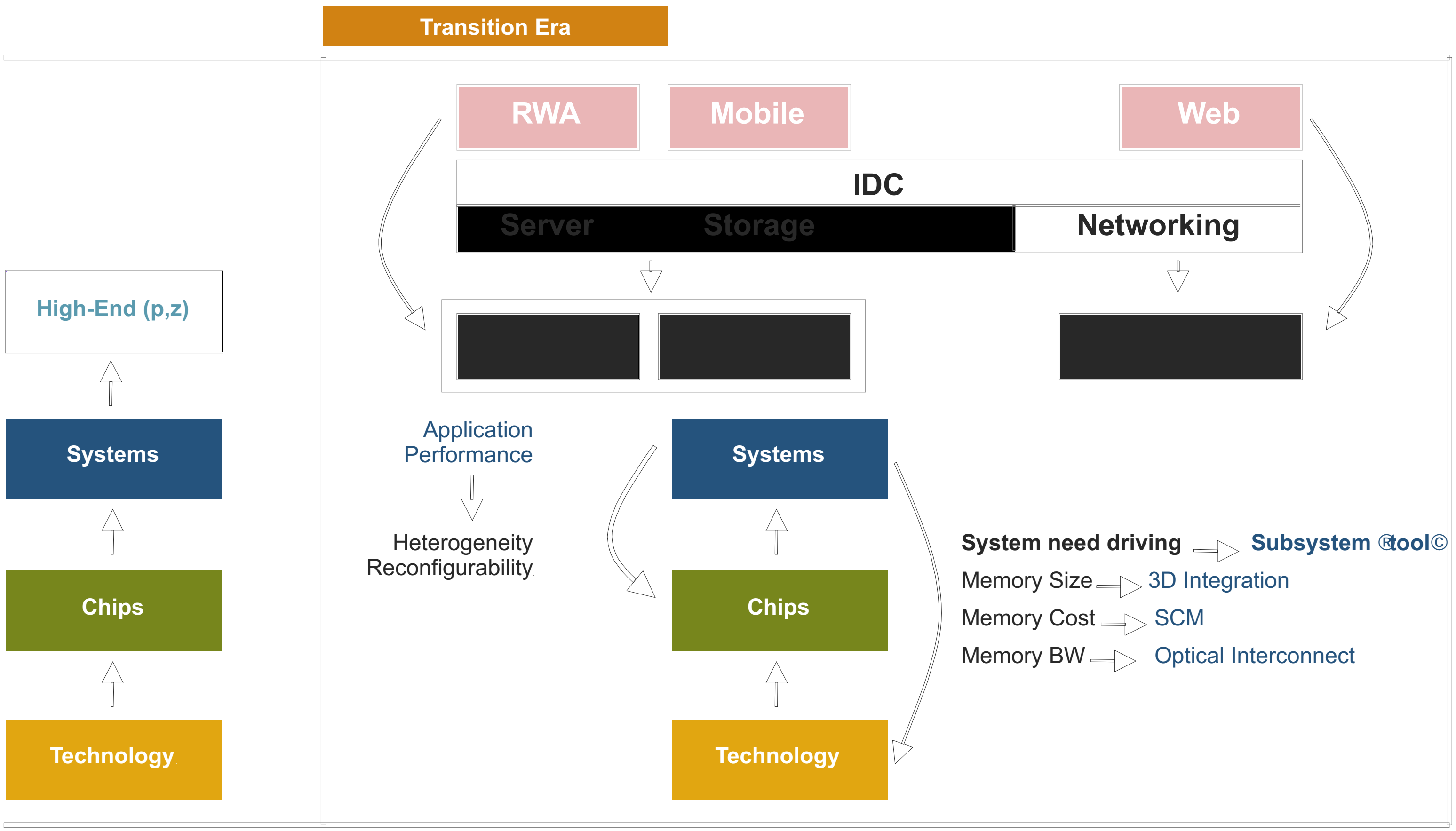
Reconfigurable Logic

- Reconfigurable logic offers substantial performance, versatility and power consumption
  - At the cost of software complexity
- FPGAs are being deployed in general-purpose systems and appliances
- The multi-core transition opens a window of opportunity for reconfigurable logic
  - Multicore also requires software modifications/rewrite
- Impact to architecture
  - What is the most effective architecture of a system with reconfigurable logic?



Increasing Reliance on System and Software for Performance

- **Leverage point is moving up the stack**
  - **Driving performance and cost requirements back to technology and chips**
- **Significant interdependence of technology with chips, systems and software**



The CMOS Transition Era

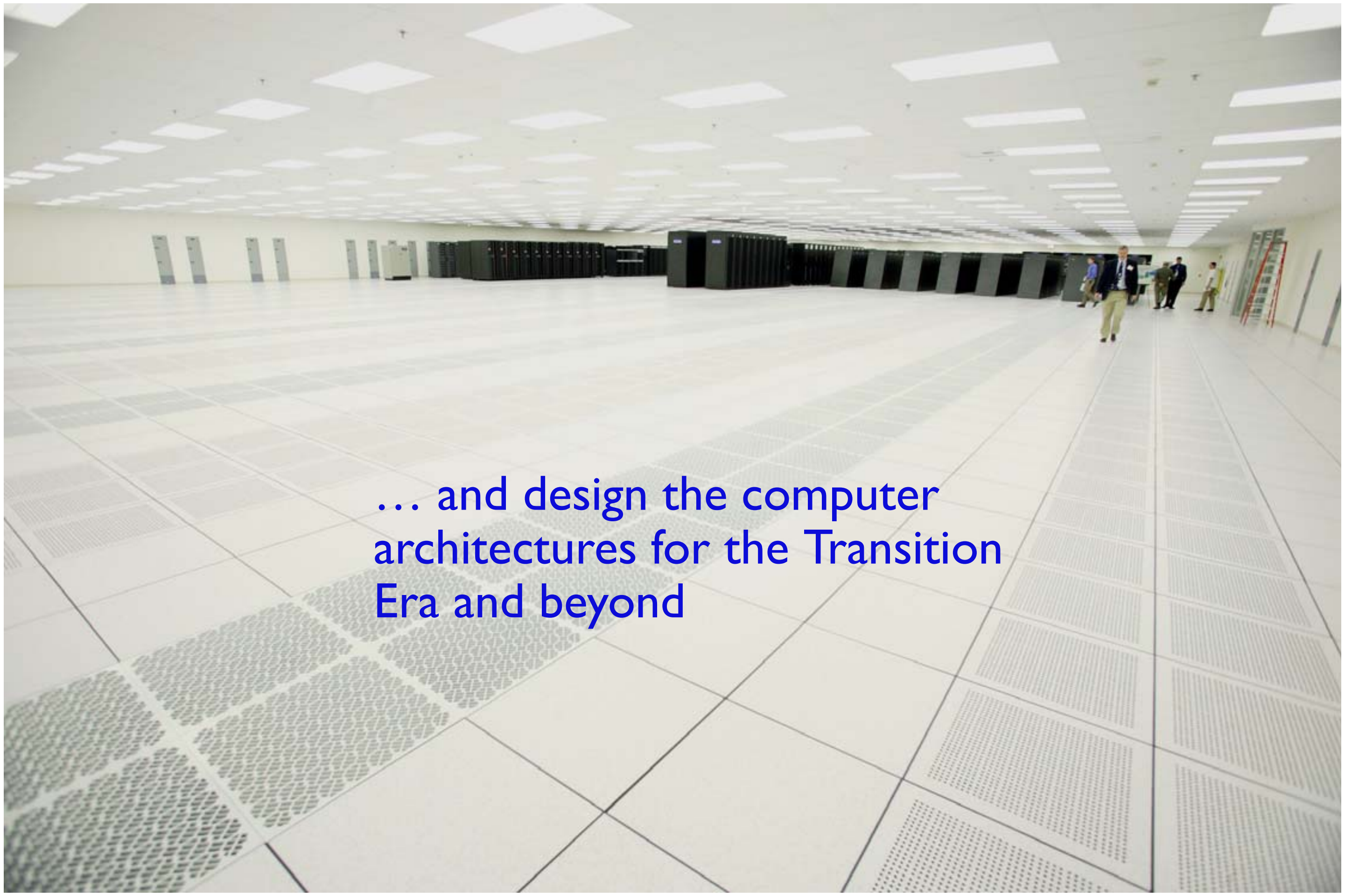
- Power continues to be the principal concern facing computer architectures today
- Not discussed today but also gaining relevance: Reliability
  - Design, manufacturing and operational challenges to overcome defects and failures
- Entering the "CMOS Transition" Era in Computer Architecture
  - Technologies other than CMOS will be deployed to continue historic performance growth trends
    - 3D silicon integration, phase-change memory, optics on chip, reconfigurable logic
  - Packaging technology will grow in importance, as chips increase in complexity
  - Systems based on chips with multiple processors (cores) are becoming widespread.
    - Generally, they are characterized by improving throughput performance and power-performance, but not single-thread performance
    - Innovation needed to improve single-thread performance
      - Programming languages, compilers, tools, etc.
- Parallelism levels that once were only within the high performance computing (HPC) domain are becoming more common, and will need to be exploited at all levels of the software stack

Where are we going?

- In about 10 years, there will be about 50B transistors in a single chip
  - What do we do with these many transistors? What are the best ways to exploit them?
  - How do we design chips of this complexity?
- Entering eras beyond Giga
  - Teraflops, Terabytes, etc.…   "Era of Tera" (Intel)
- Exascale systems in 10-20 years
  - 1000x what we have today
  - Roadrunner: 1 Petaflop today -- 1 Exaflop in late '10s
- Data being generated at increasing rates and at increasing speeds
  - Human generated, machine generated …….
  - No longer possible to store all the data - processing on the fly (stream processing)

- These are just some examples ….
  - All of them will lead to advances in Computer Architecture

Understand the Technology Trends …



IBM Research

August 2008