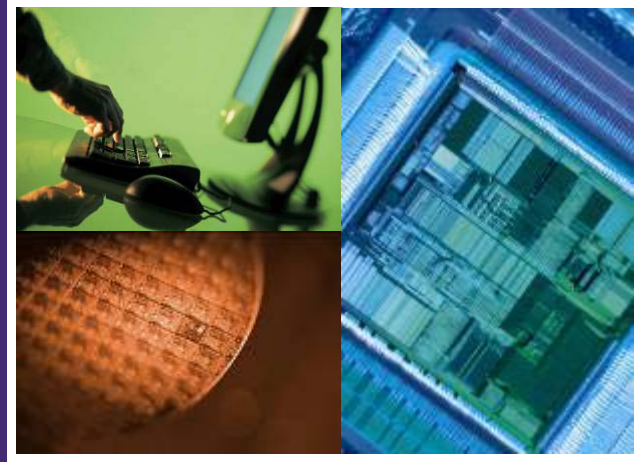
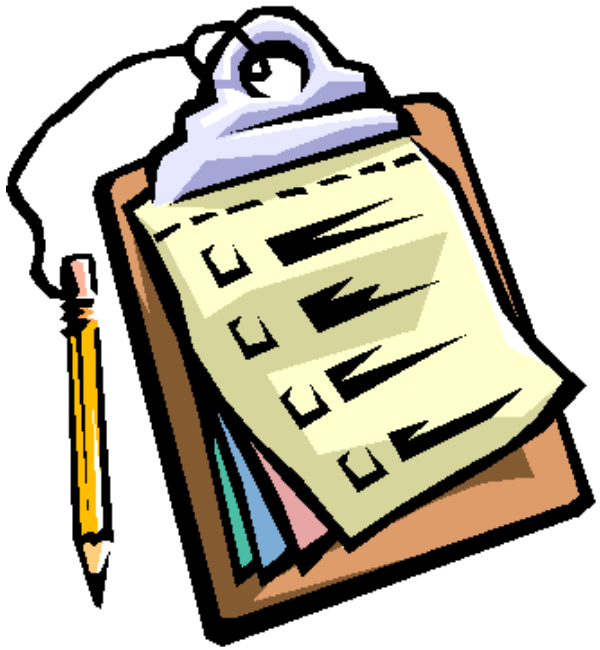
Speeding EDA Design Cycles with HPC technology

Glenn Newell
Sr. Staff IT Architect
Synopsys



SYNOPSYS®
Predictable Success

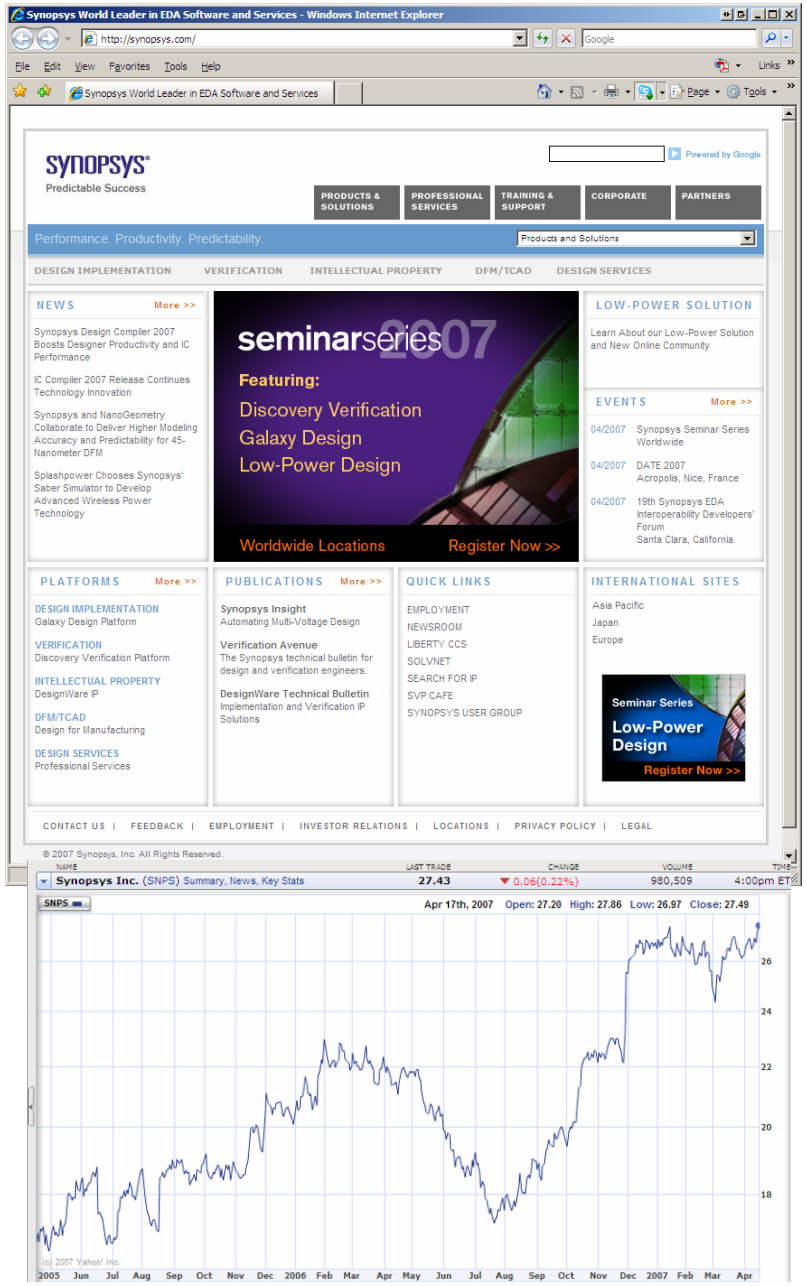
Agenda



- Synopsys Intro
- IT Infrastructure
- Problem Space Overview
- Enterprise Class Infrastructure
  - IT challenges
- High Performance Computing (HPC)
  - Solutions
- Results
- Best Practices
- Getting The Message Out

Synopsys

- “A world leader in semiconductor design software”
- Company Founded:1986
- Revenue for FY 2006: $1.096 billion
- Employees for FY 2006: ~5,100
- Headquarters: Mountain View, California
- Locations: More than 60 sales, support and R&D offices worldwide in North America, Europe, Japan, the Pacific Rim and Israel



SYNOPSYS®
Predictable Success

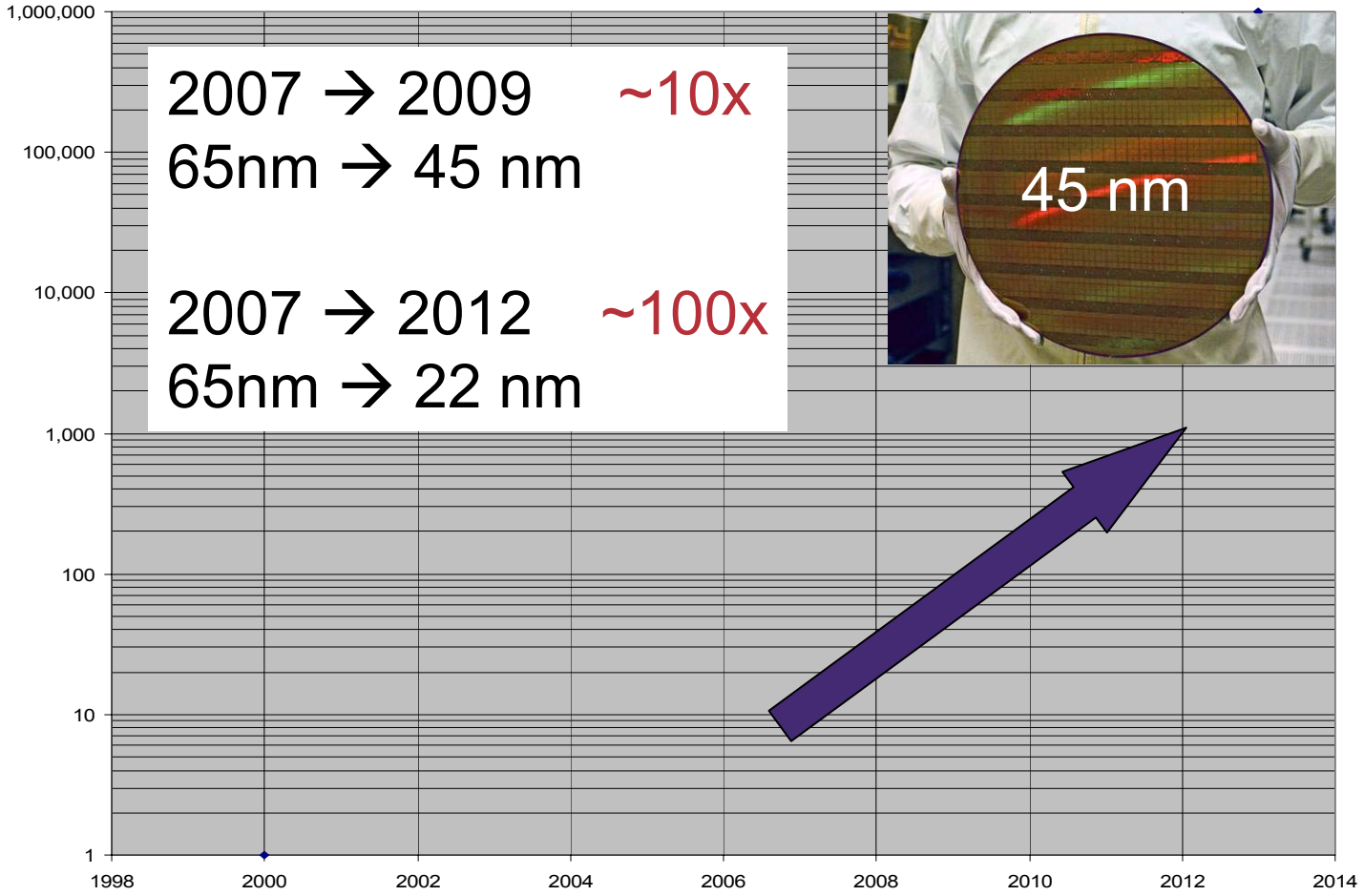
Synopsys IT (2007)

- Over 60 Offices World Wide
- Major Data Centers
  - 5 at HQ
  - Hillsboro OR
  - Austin TX
  - Durham NC
  - Nepean Canada
  - Munich Germany
  - Hyderabad India
  - Yerevan Armenia
  - Shanghai China
  - Tokyo Japan
  - Taipei Taiwan
- 2 Petabytes of NFS Storage
- ~15000 compute servers
  - 10000 Linux
  - 4000 Solaris
  - 700 HPUX
  - 300 AIX
- GRID Farms
  - 65 farms composed of 7000 machines
  - 75% SGE 25% LSF
- Interconnect
  - GigE storage
  - Fast E clients

- #242 on Nov. '06 Top400.org
- 3.782 TFlops on 1200 Processors

Problem Space

Estimated EDA Relative CPU Cycles Required



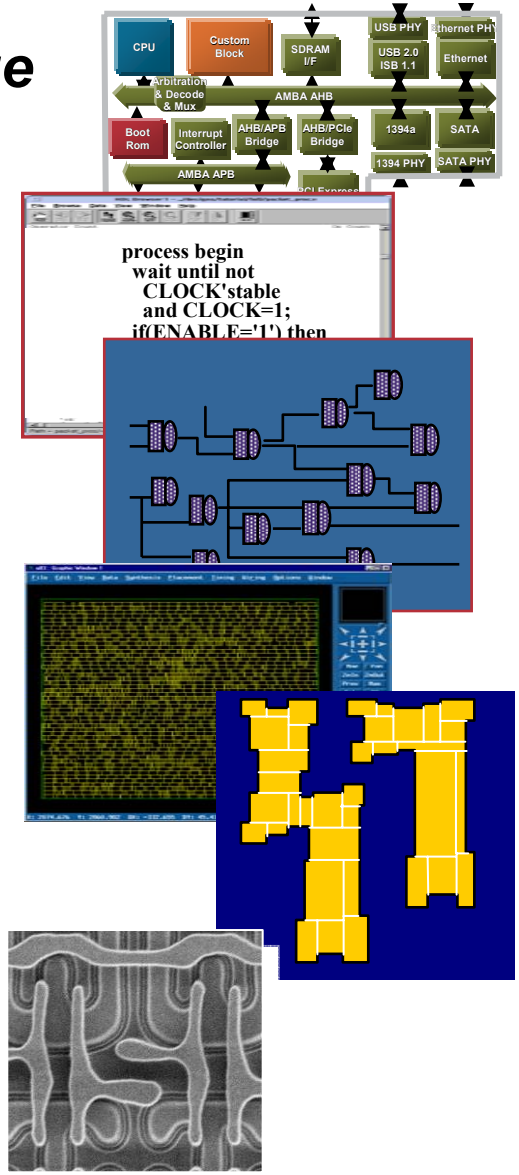
SYNOPSYS®
Predictable Success

Problem Space

Designing a (*Large*) Chip

Complexity and Miniaturization Continue

Size	Units	Level	Who is affected
10^3	Blocks, IP	System	Designer
10^5	RTL lines	RTL	Designer
10^8	Gates, Bits	Netlist	Synthesis
10^9	Transistors	Circuit	Cells / Memory / Analog
10^{12}	Polygons	Layout	Place & Route
10^{13}	Trapezoids	Mask	Mask Synthesis



* CMOS, mostly digital, 65nm, >200mm^2

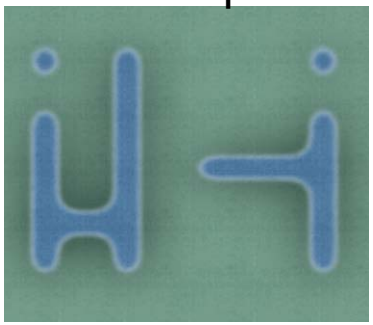
SYNOPSYS®
Predictable Success

Problem Space

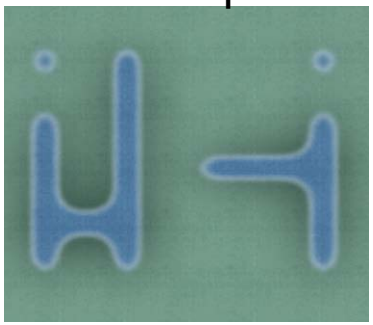
No Image = No Yield = No Product = No $$$$

Original Layout Pattern

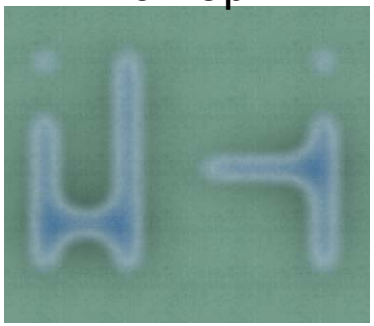
0.25µ



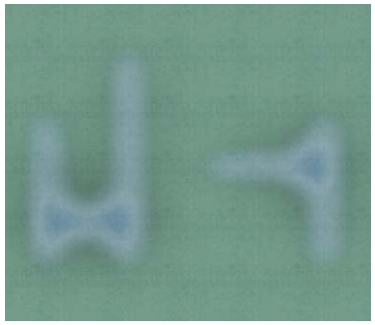
0.18µ



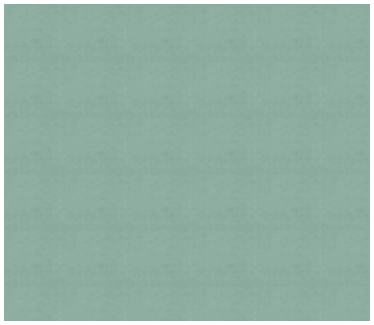
0.13µ



90nm

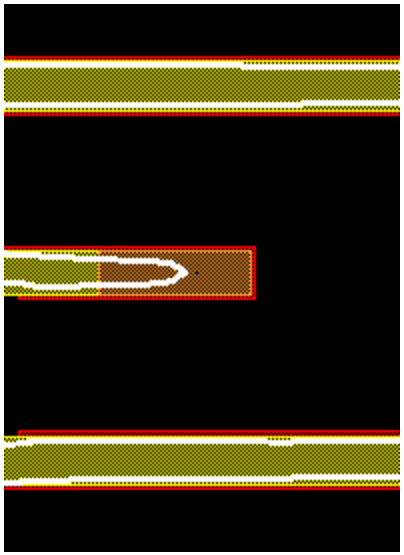


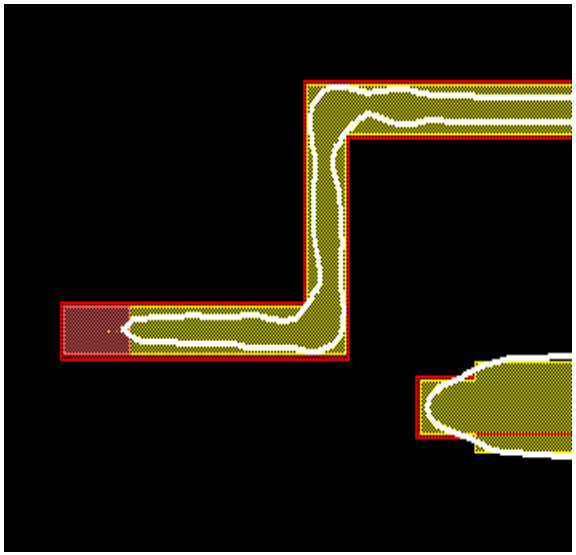
65nm

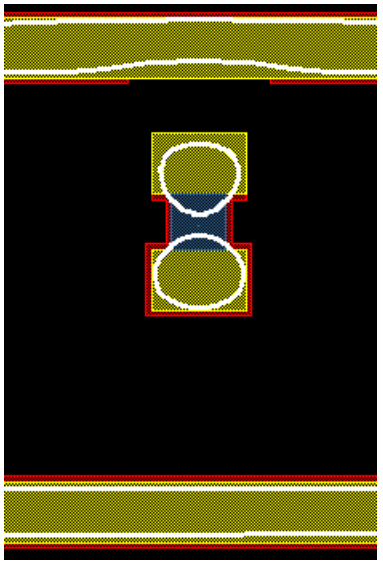


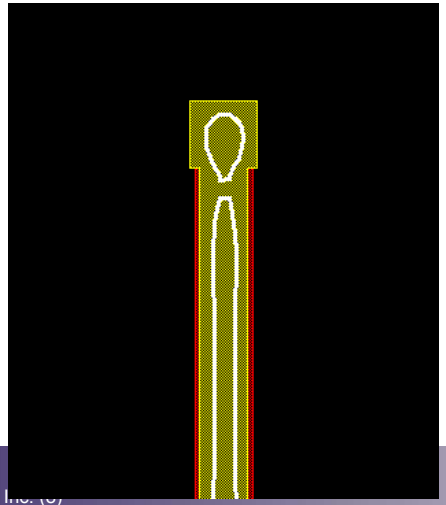
Problem Space

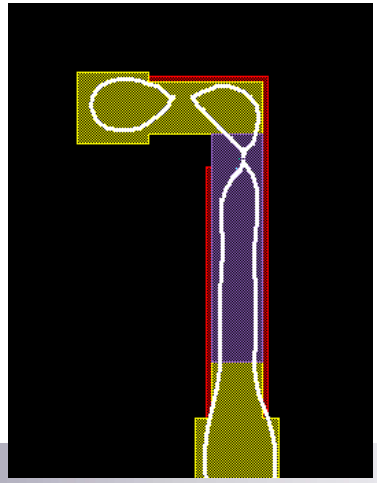
65nm Hot Spots Examples

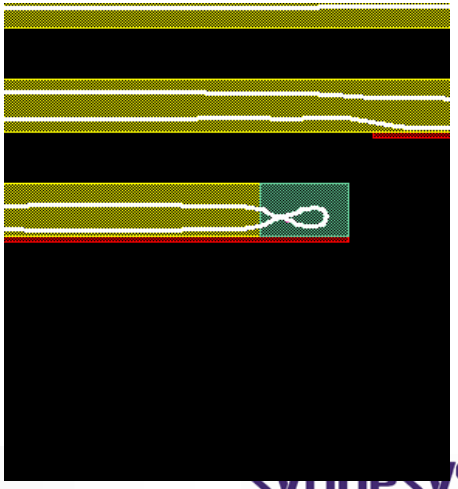








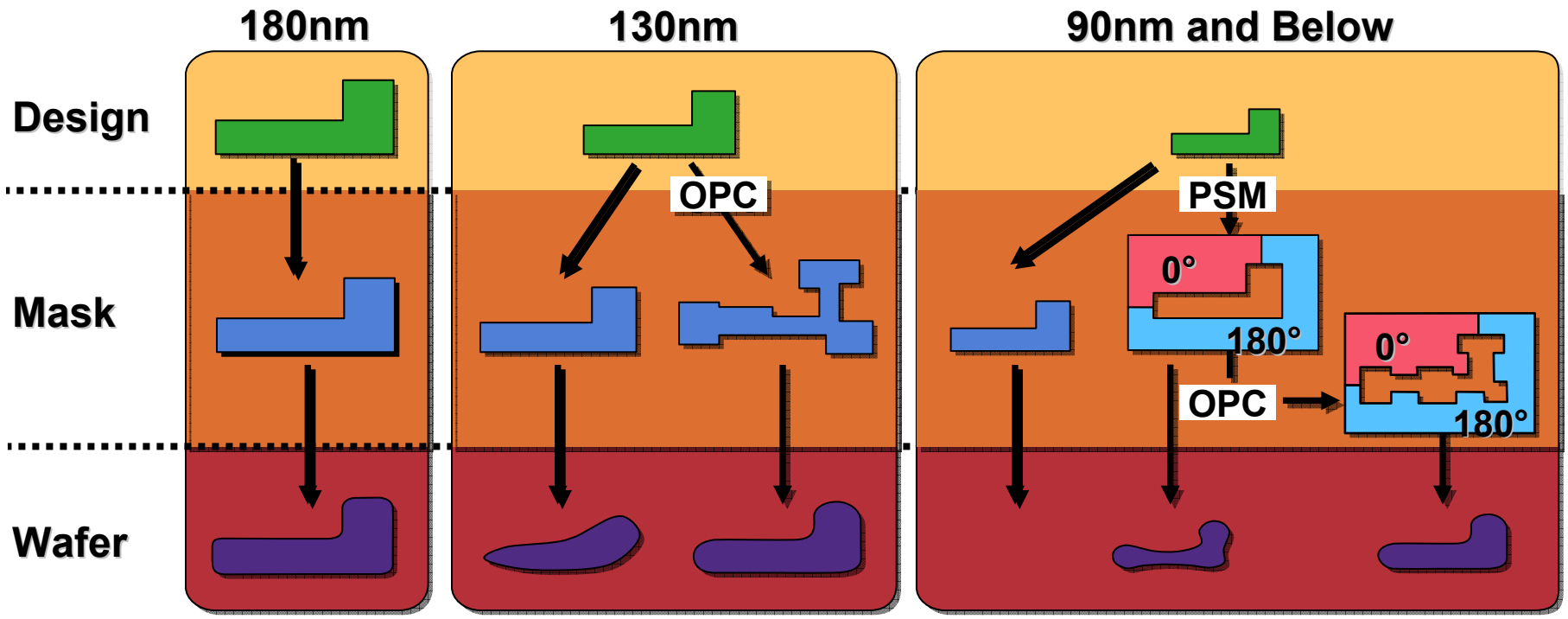




Problem Space

Chip Manufacturing

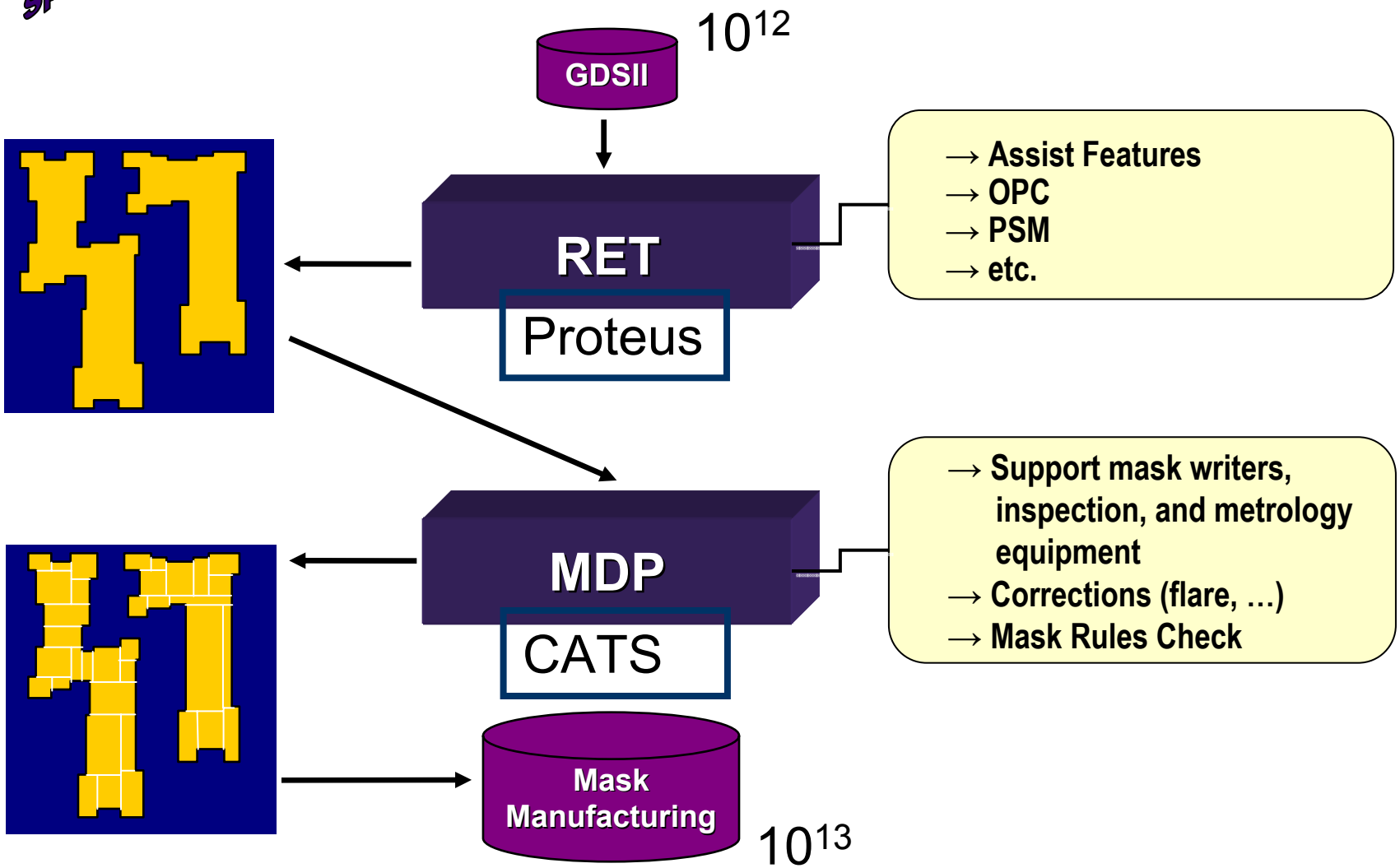
Design & Mask "Enhancements" Become the Norm



SYNOPSYS®
Predictable Success

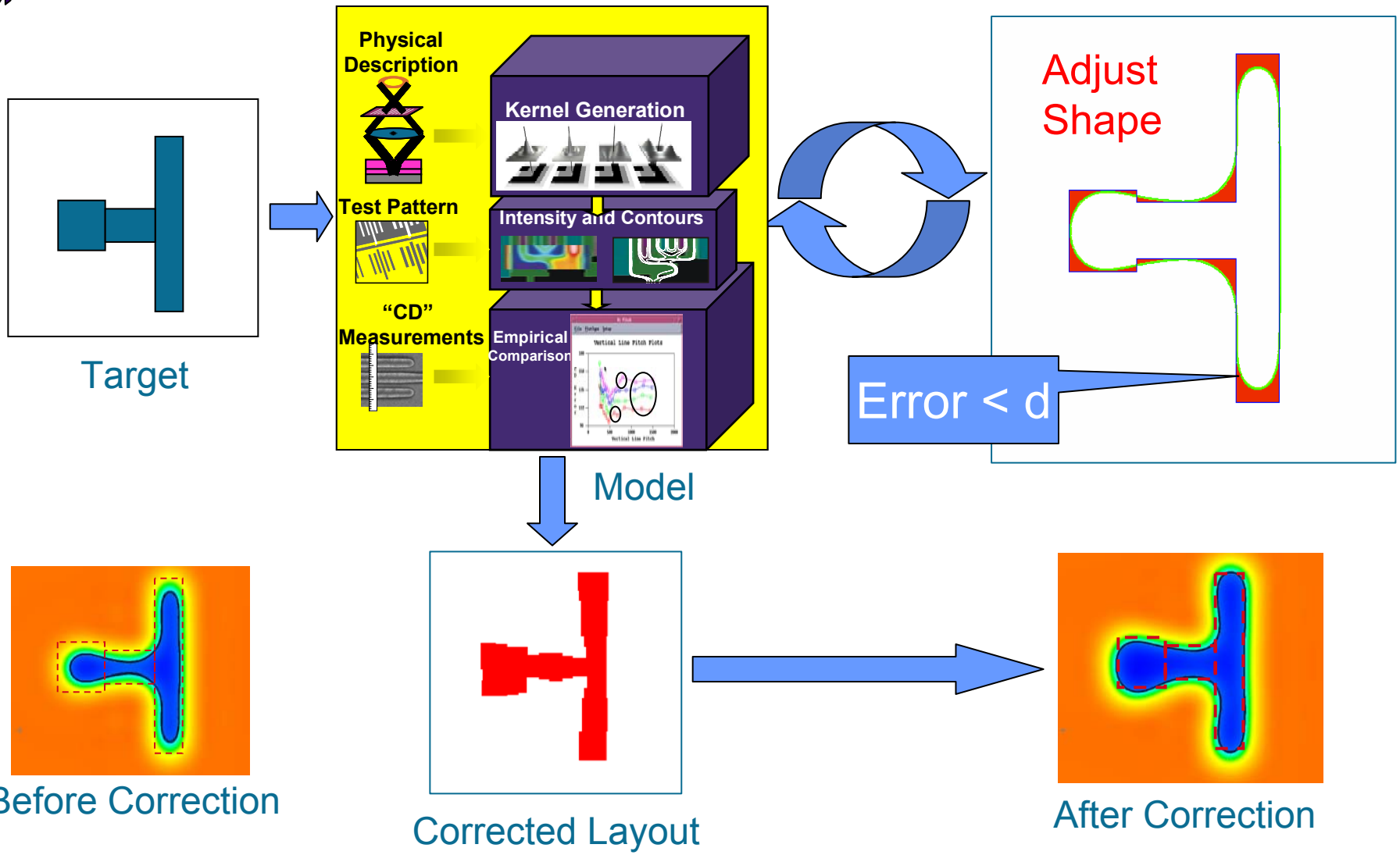
Problem Space

Mask Synthesis Flow



Problem Space

Model-Based (Simulation-Based) OPC



Problem Space

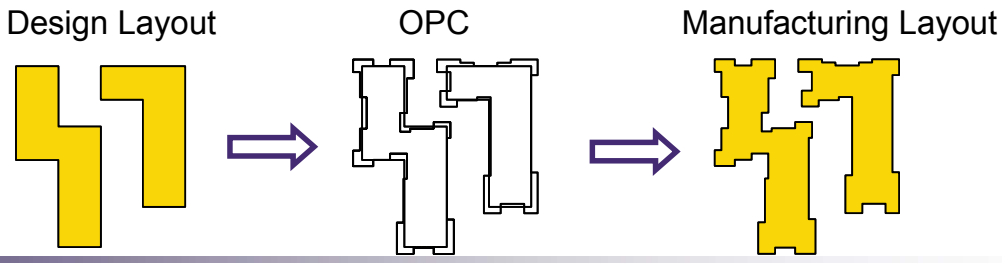
Computing and Chip Manufacturing

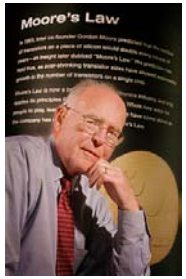
Sequential Processing – Not the answer!

An Example: OPC simulation using sequential processing

$$8\,iterations \times 3\times 10^{4}\left[\frac{instruction}{polygon}\right]\times 10^{12}\left[polygons\right]\times 3\times 10^{-9}\left[\frac{\text{sec}}{instruction}\right]$$

~ 10^{9}[sec] ~ 1000 days



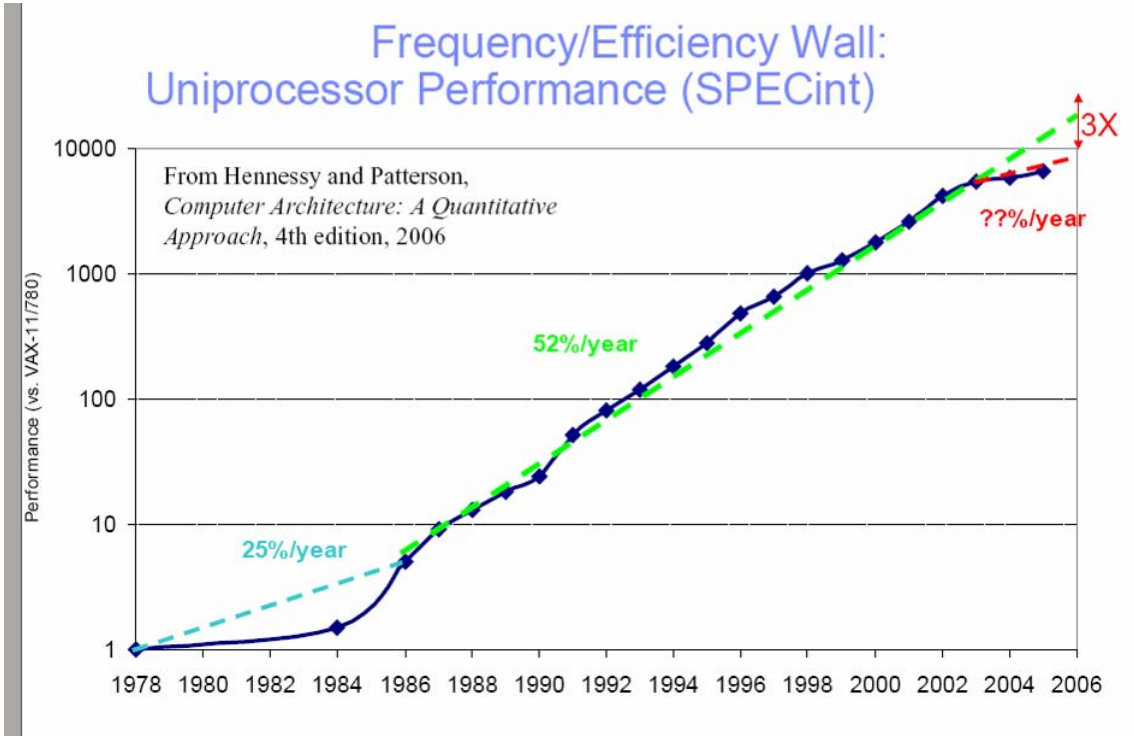


A Glitch in Moore's Law?

Single thread performance hitting limits

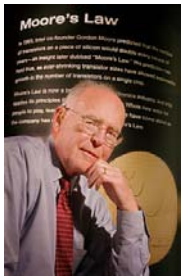
Problem Space

Moore's "Law" states that the performance of semiconductor technology will grow exponentially over time. Exponential growth has a constantly increasing rate of change. However, recently there has been a significant decline in the rate of increase of CPU performance, which indicates a deviation from Moore's Law



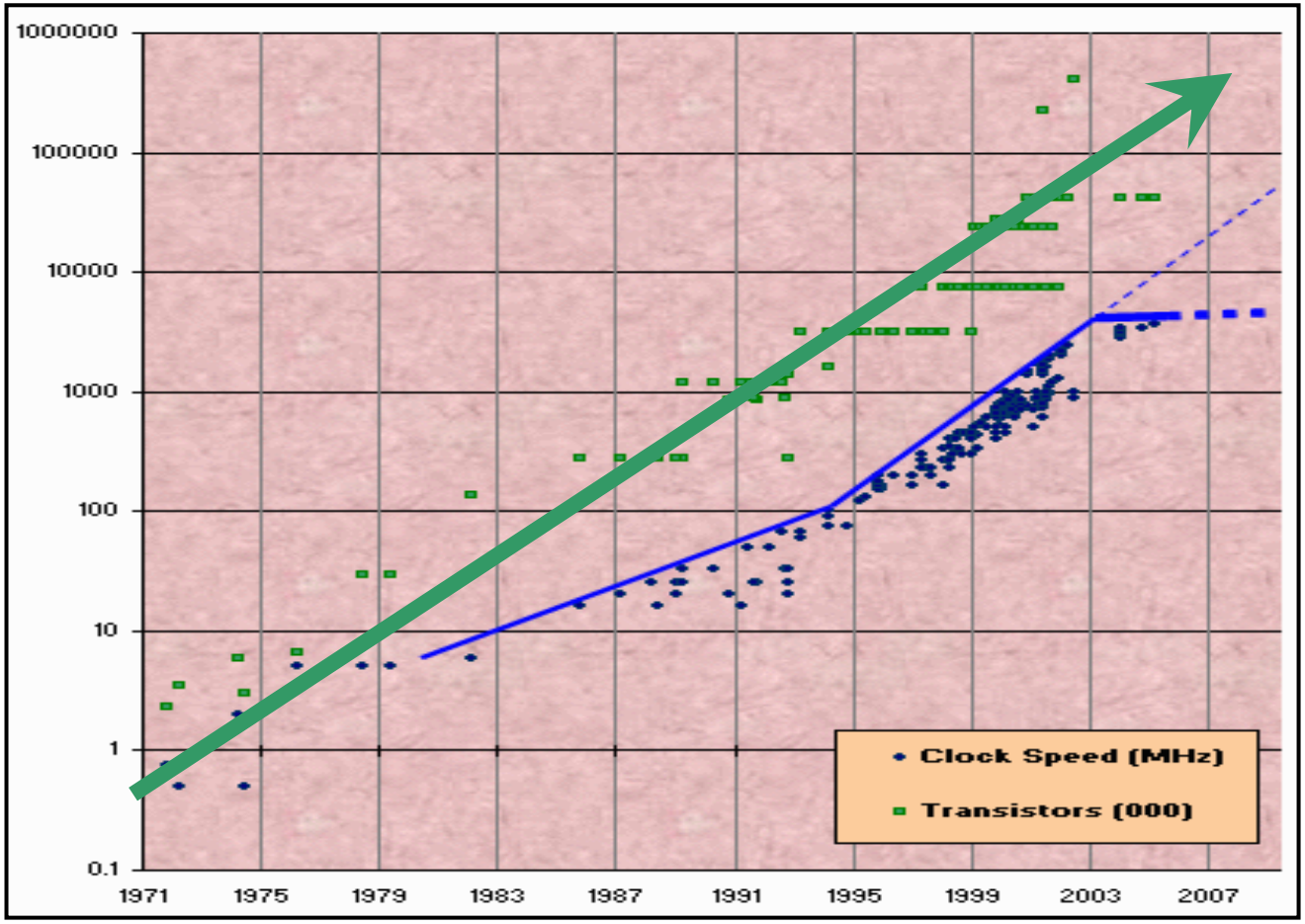
Textbook Figure showing 3x gap in actual single core performance vs. projection starting in 2003 until the present.

- SPECint is a computer benchmark specification for CPU's integer processing power. It is maintained by the Standard Performance Evaluation Corporation (SPEC).
- Floating point performance graph would like similar.



CPU: Single thread performance hitting limits

Problem Space

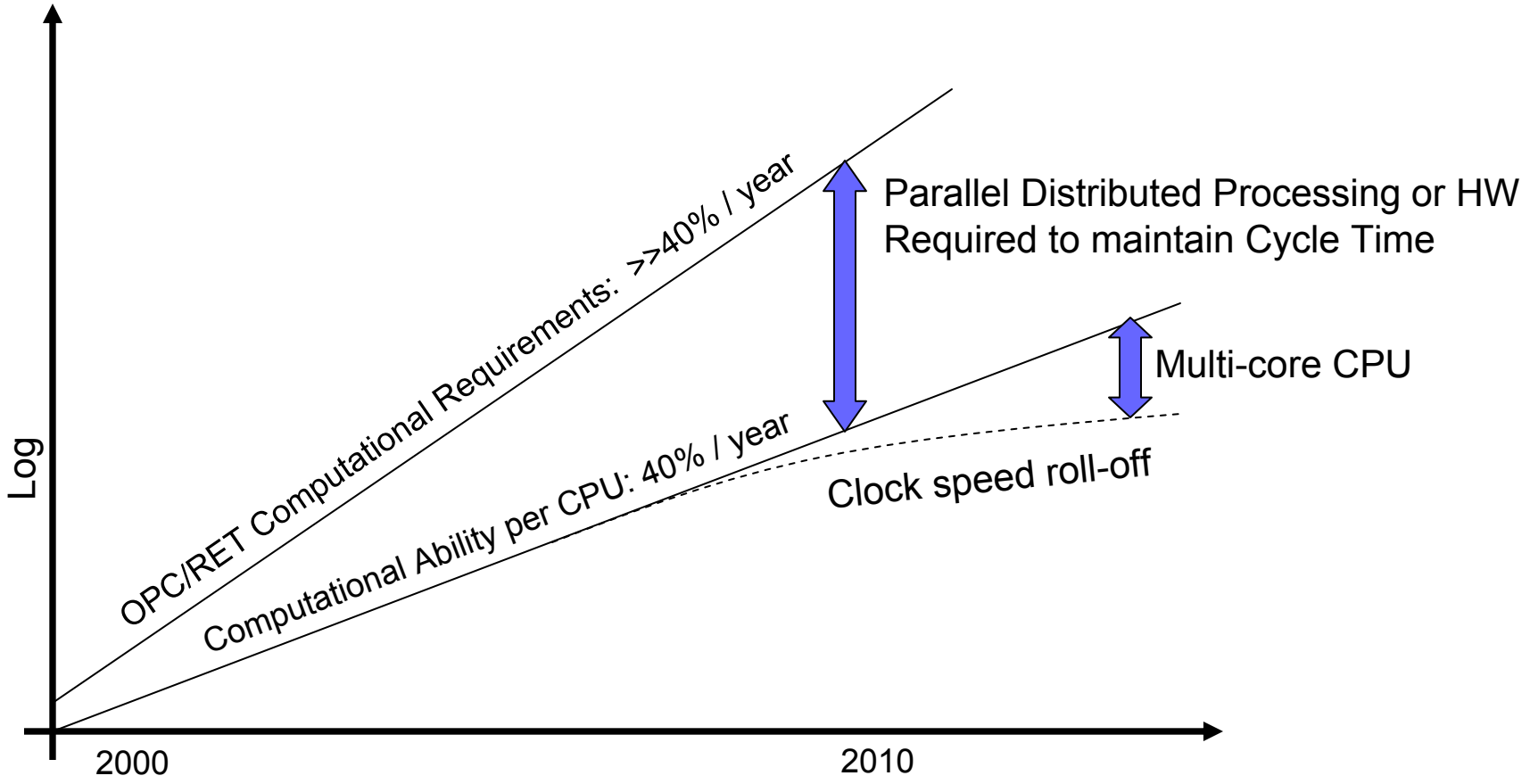


Moore's Law is alive and well. Clock Speed hits limits.

Problem Space

Computing & Chip Manufacturing

Work demand increasing faster than CPU capabilities

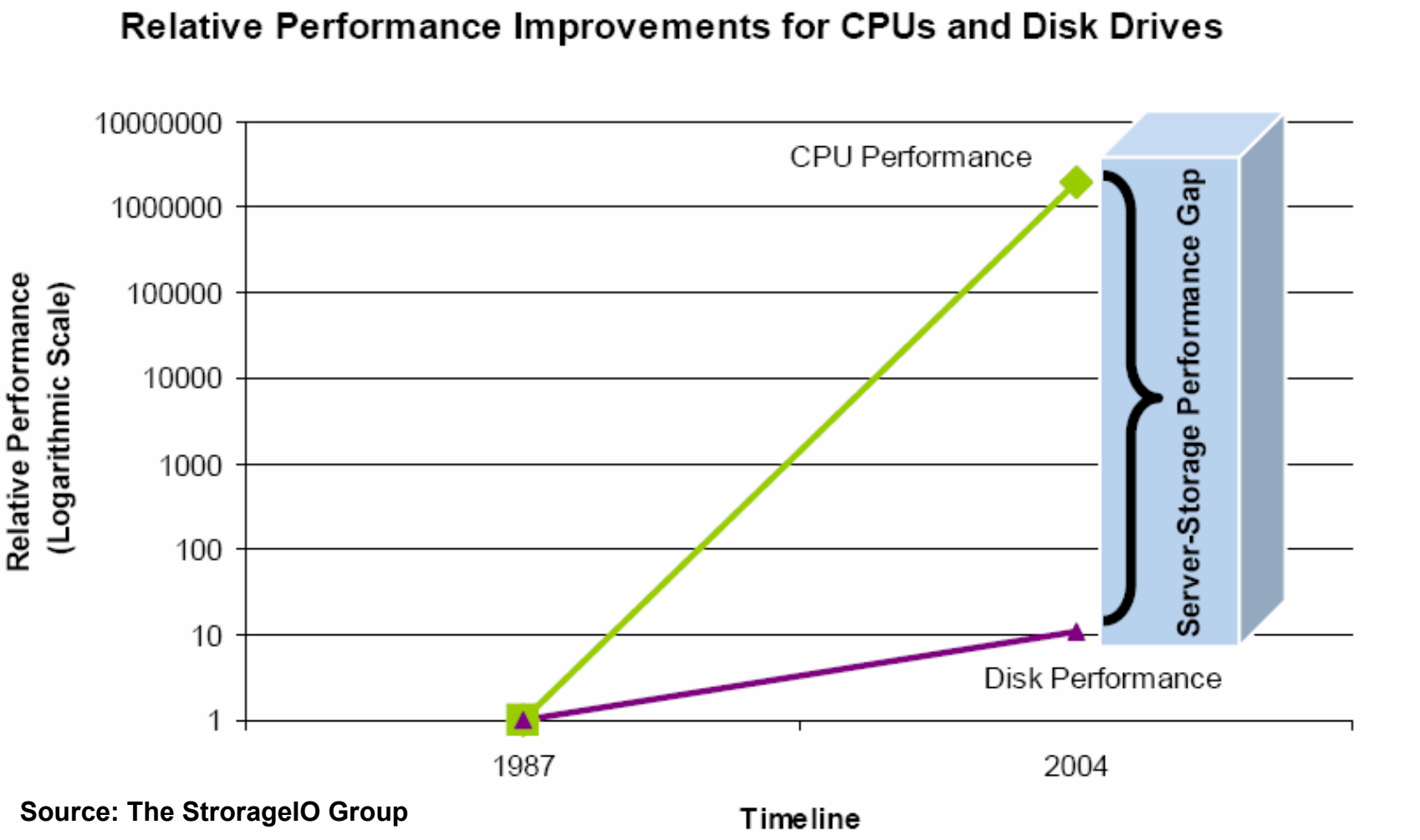


SYNOPSYS®
Predictable Success

Problem Space

I/O Performance

Server-Storage Performance Gap Increases

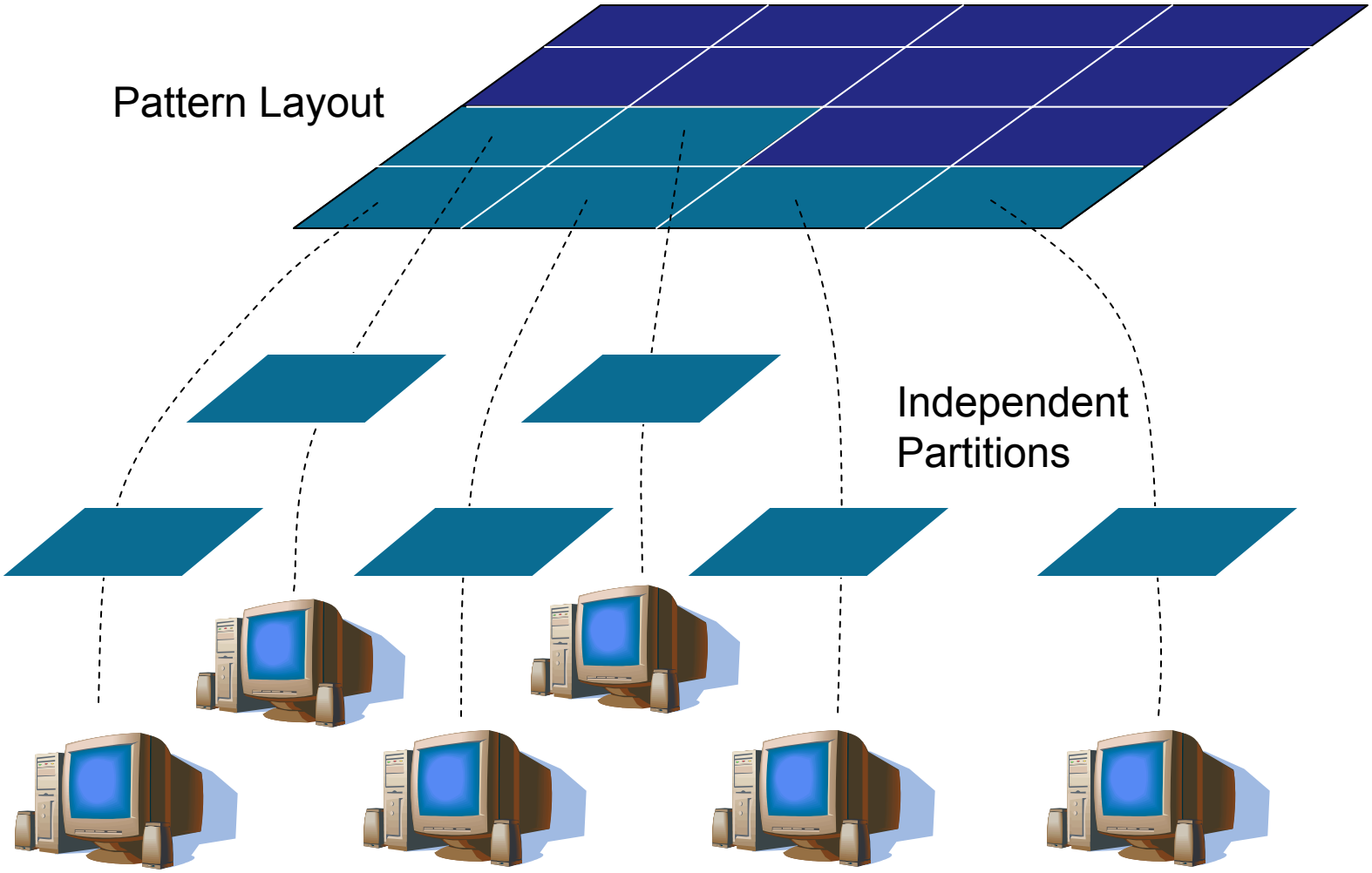


Source: The StrorageIO Group

Problem Space

Computing & Chip Manufacturing

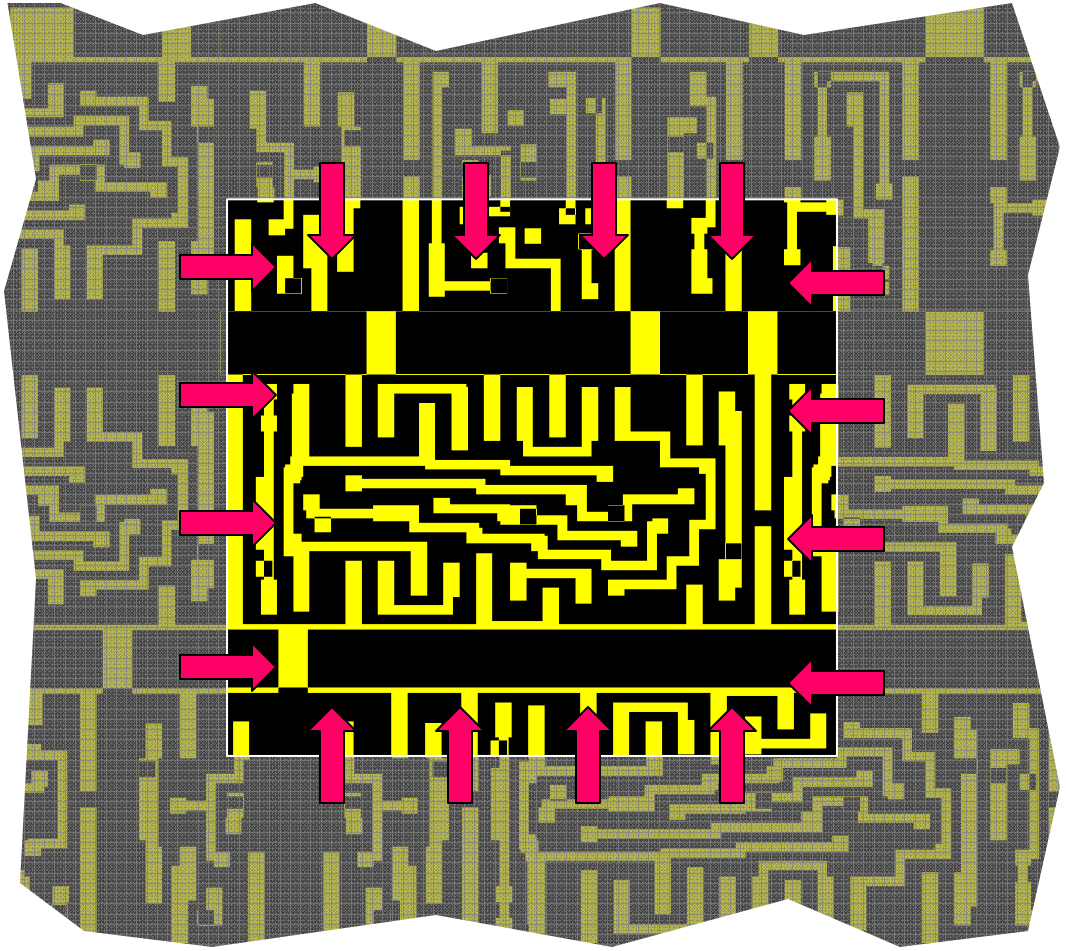
Basic Parallelization: Divide & Conquer



Model OPC & MDP
For most operations, partitions cannot be independent

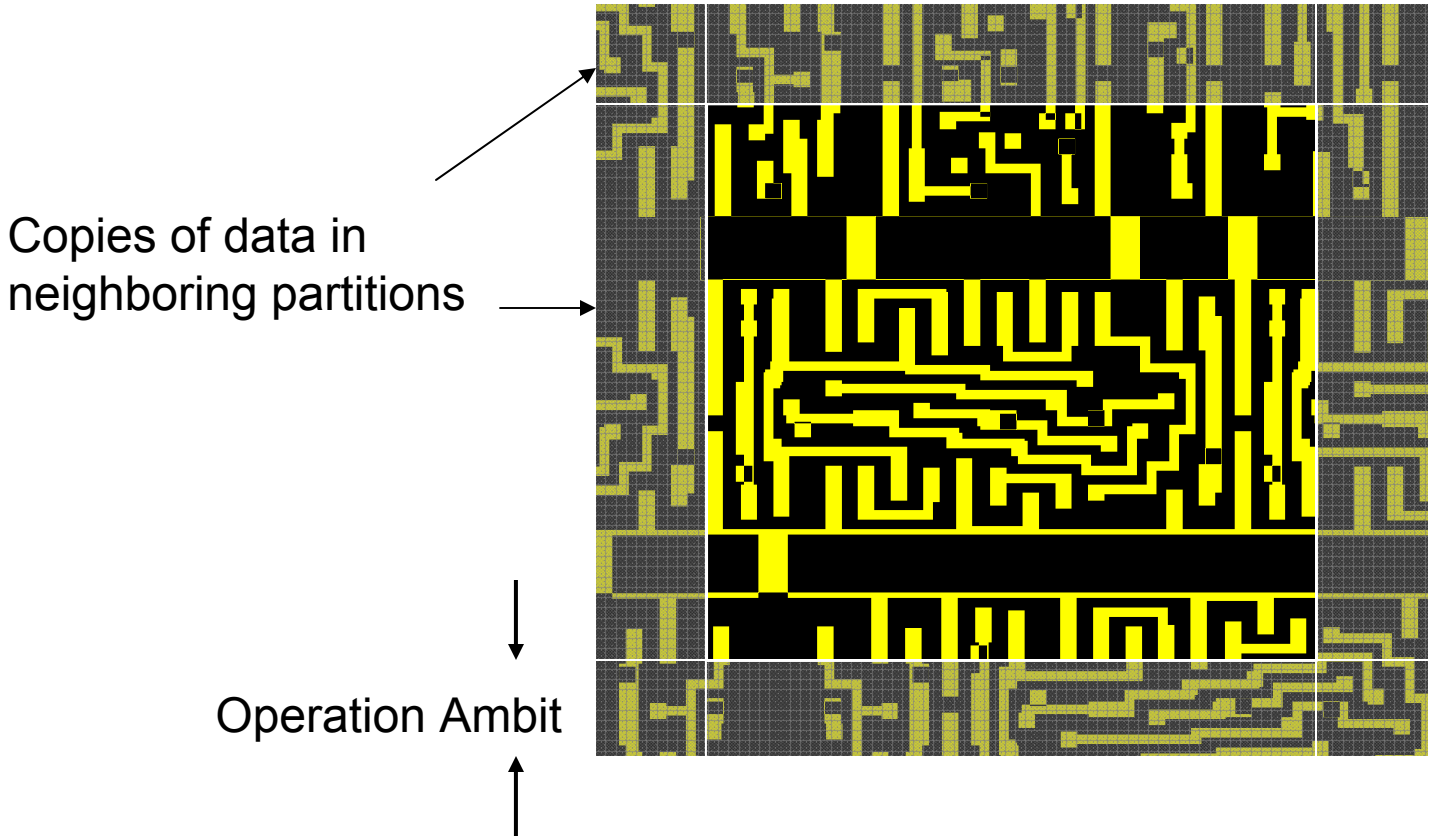
Problem Space

Feature manipulations involve inputs from neighboring features



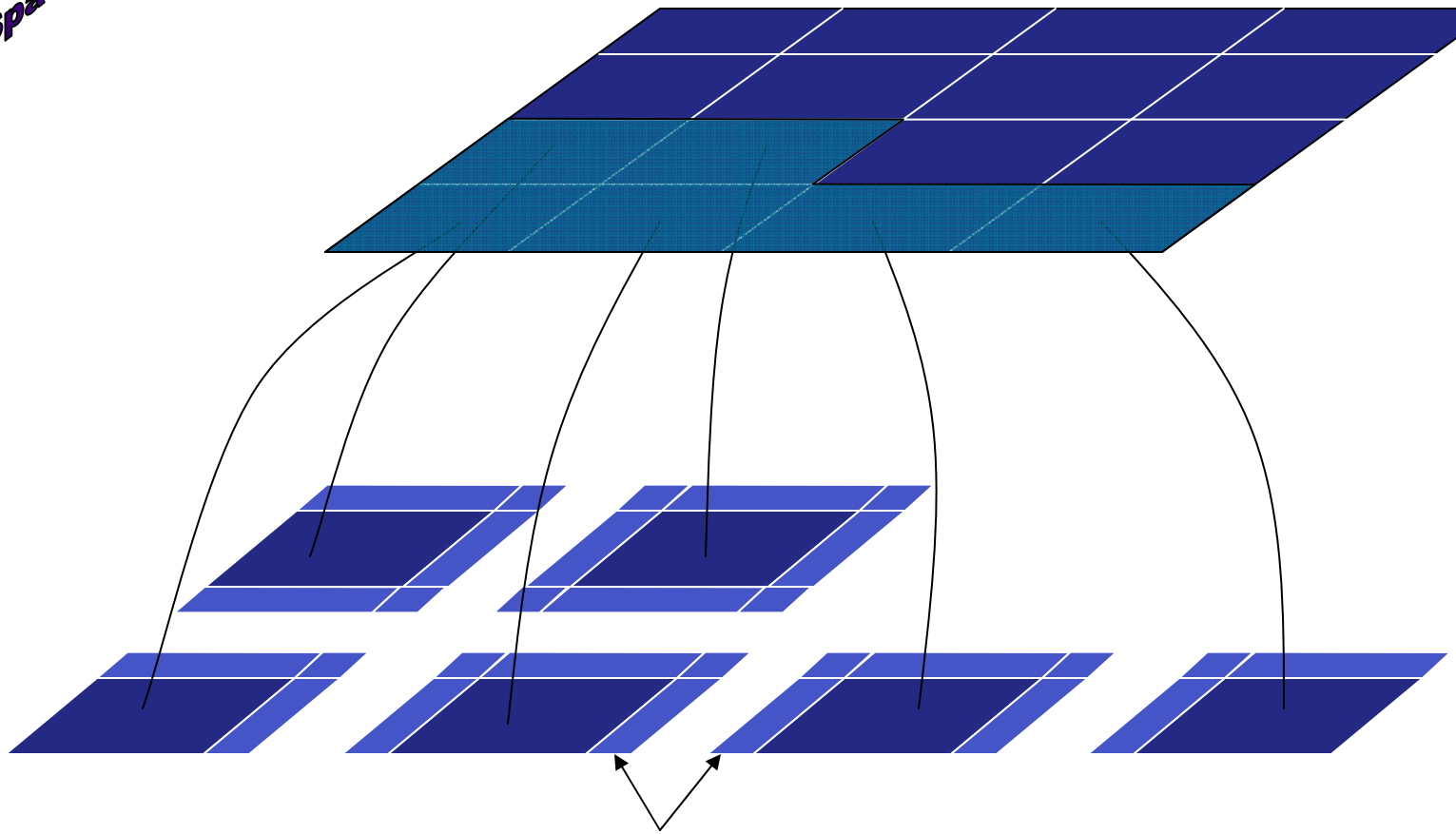
Partitions include neighborhood data

Problem Space



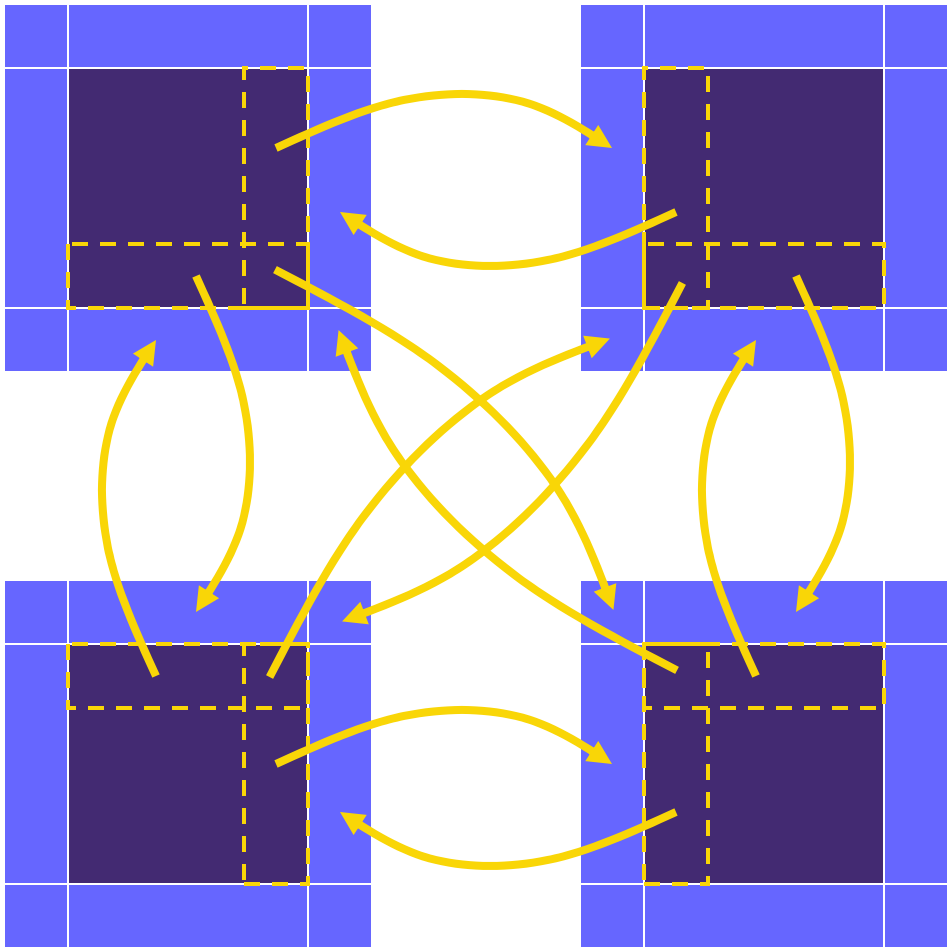
Redundant data in each partition increases total work

Problem Space



Problem Space

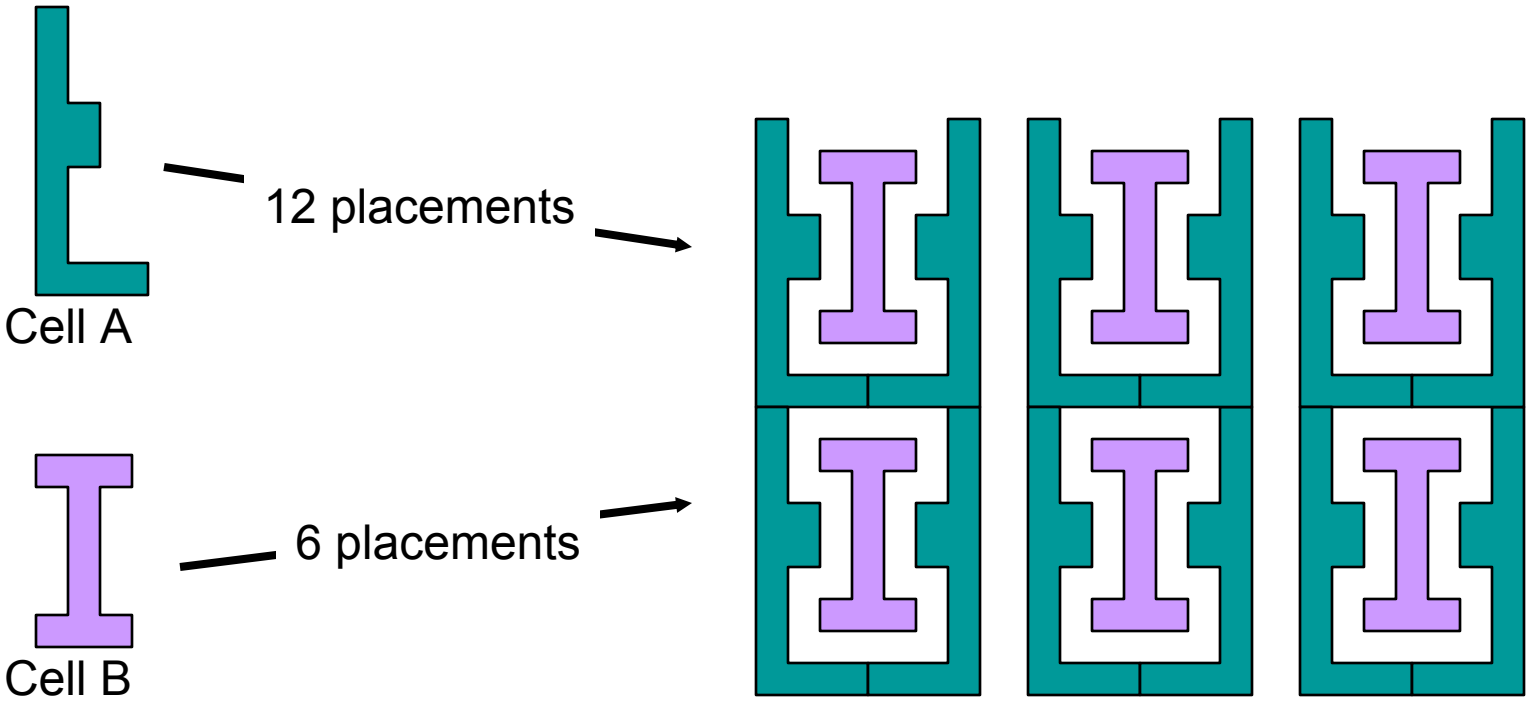
Partition size can be managed by updating results between processing stages



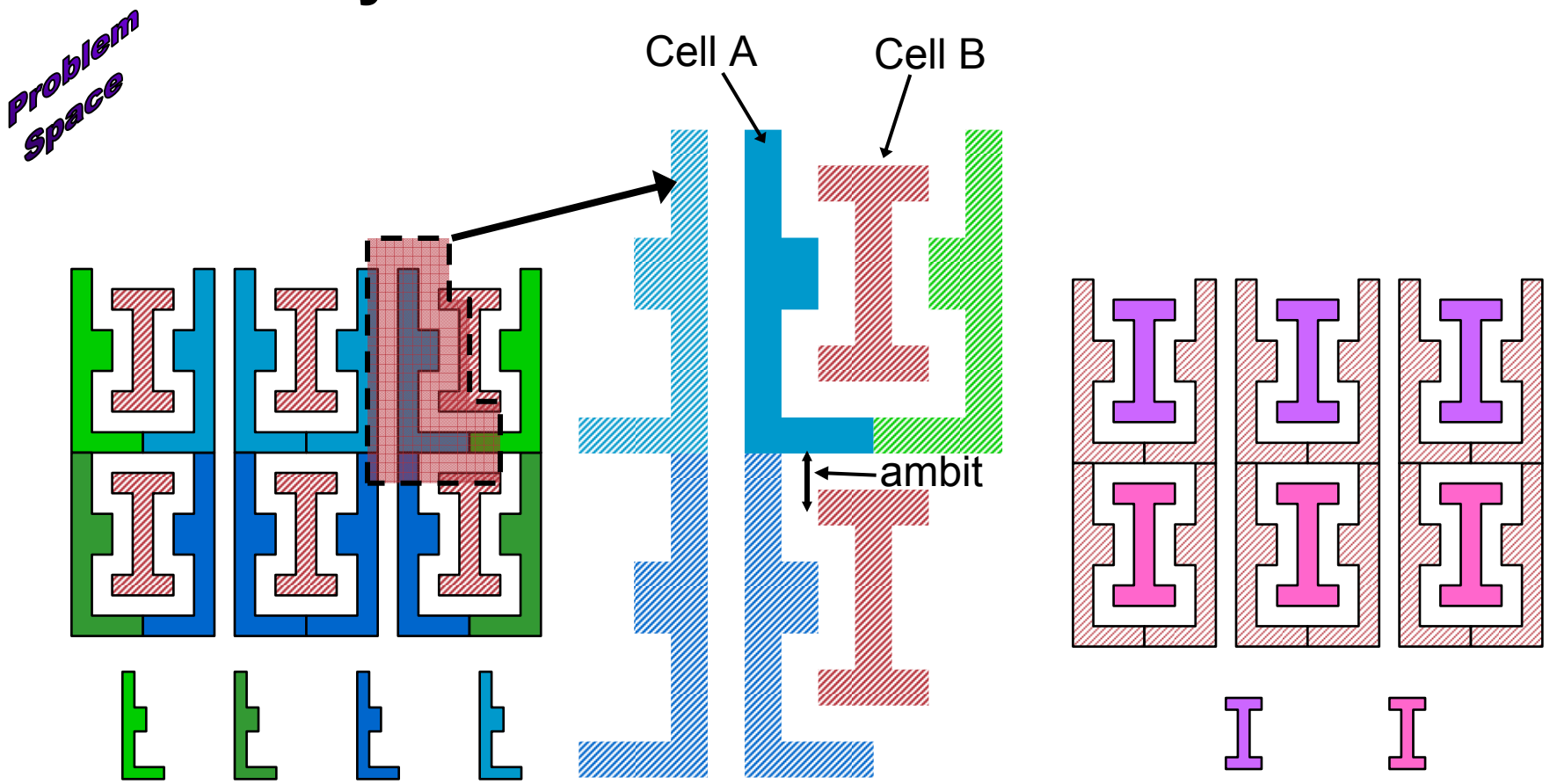
Exchange neighborhood results

Problem Space

Exploiting Hierarchy is Tricky



Each cell requires several different corrections dictated by the cell environment



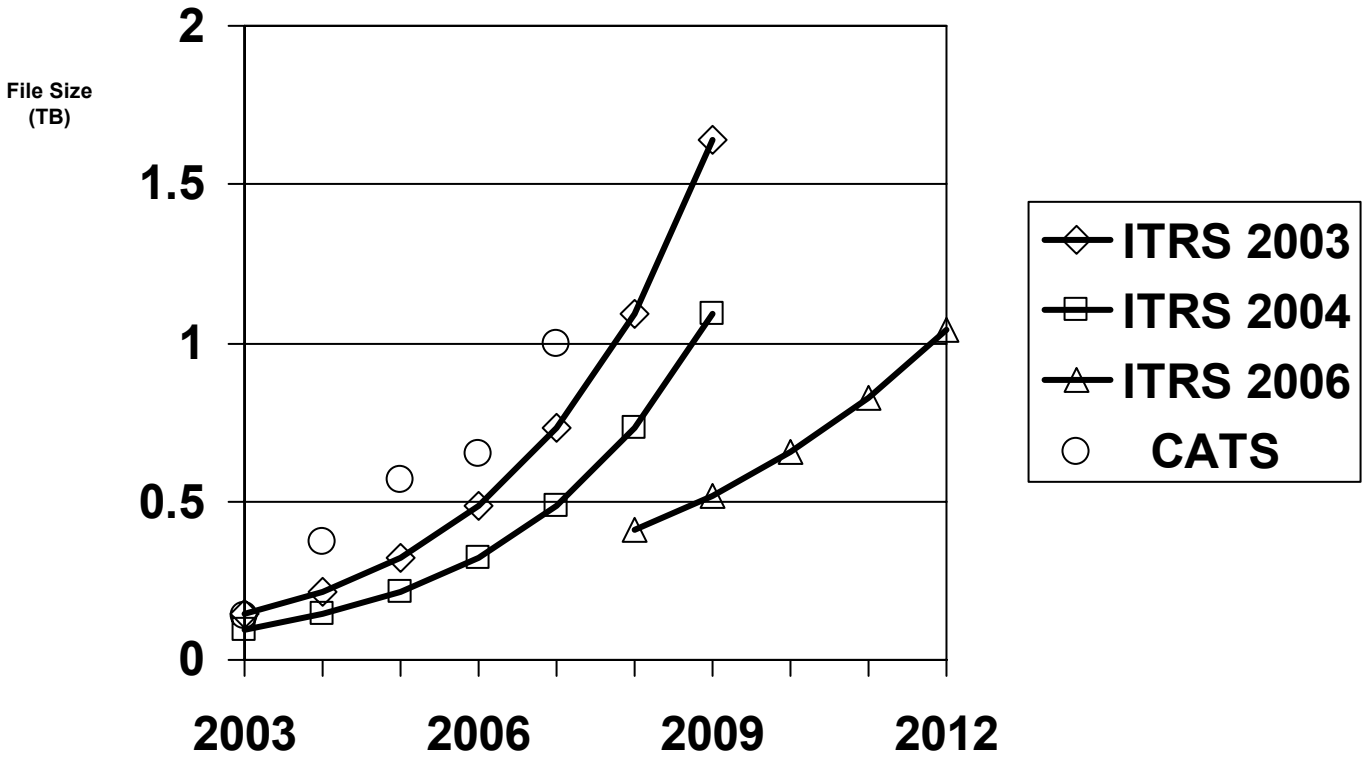
Cell A needs 4 unique corrections
(out of 12 instances)

Cell B needs 2 unique corrections
(out of 6 instances)

Drivers

Design Size

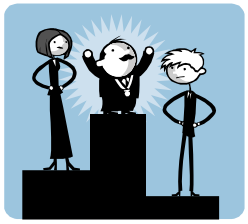
- Today, 100 GB pattern data files common
- In next few years, 1TB files will not be unusual



Drivers

The Speed of Business – Turn Around Time

- Same amount of hours in a day, despite larger chip designs

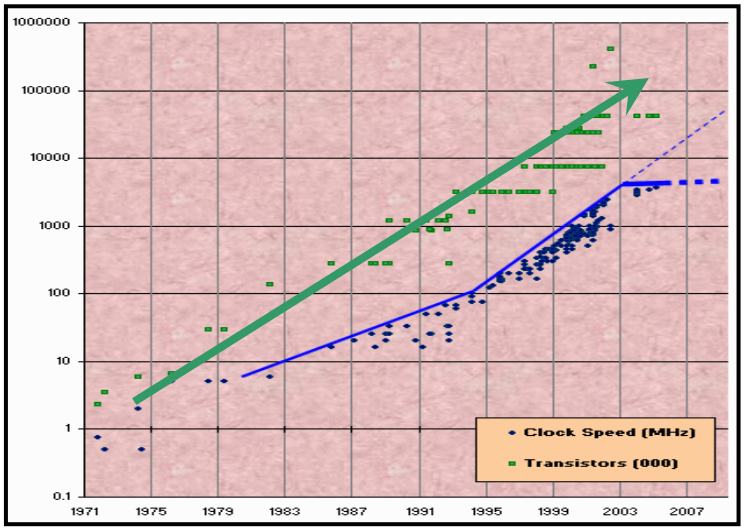


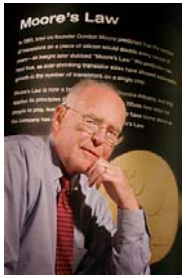
- Guaranteed Turn Around Time Drives Mask Shop Competitiveness

Drivers

Multicore processors

- The performance of individual CPUs has stopped increasing
- CPUs will be multicore from here on out
- This implies further load on interconnect and storage as individual computers become more powerful
- More Multicore effects on next slide

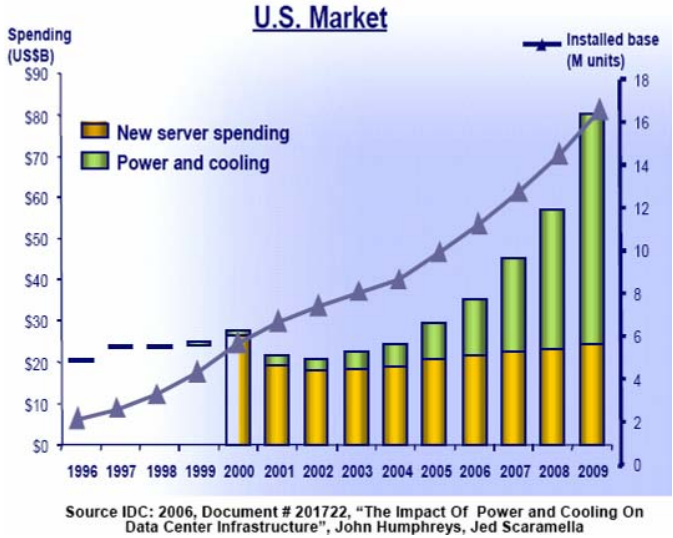




Moore's Law is alive and well. Clock Speed hits limits.

Drivers

Data Center Costs

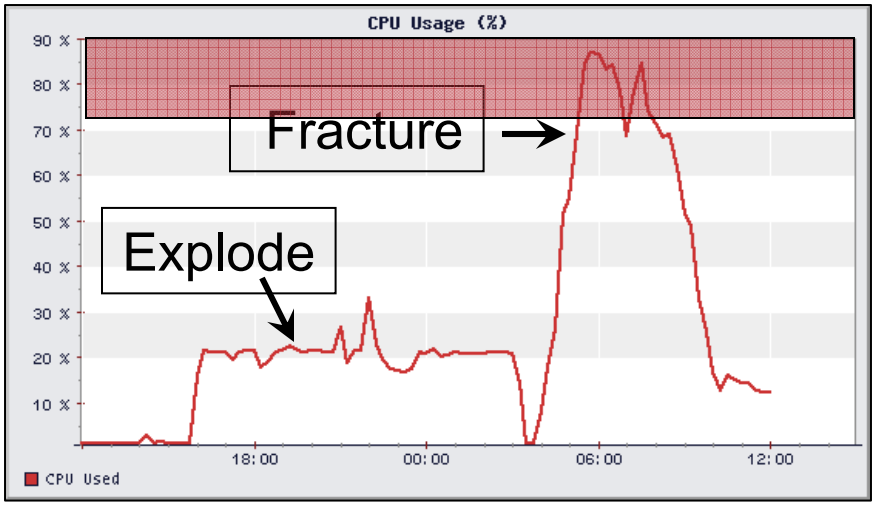


Source IDC: 2006, Document # 201722, "The Impact Of Power and Cooling On Data Center Infrastructure", John Humphreys, Jed Scaramella

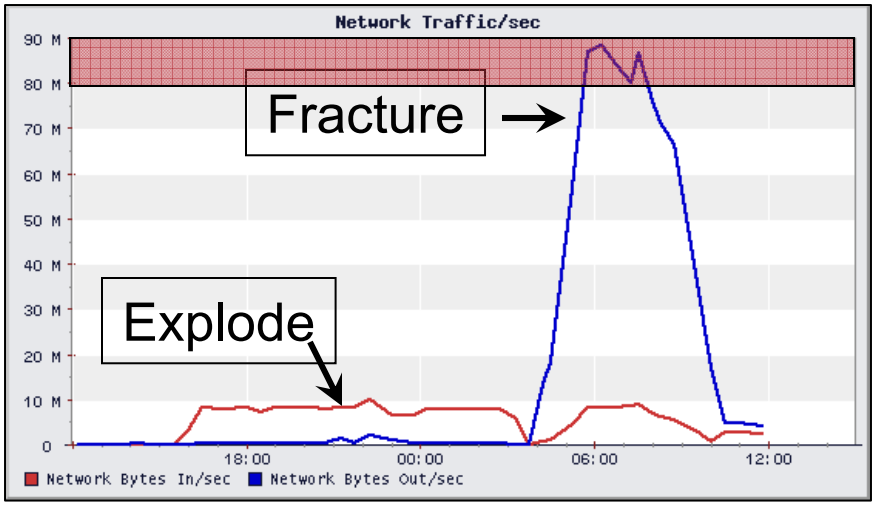
- Beginning in 2007, the Data Center Costs to house, power, and cool compute servers is exceeding their capitol acquisition cost.
- Minimize cost by using fewest CPUs + most efficient IT infrastructure for required TAT
- More money spent up front (CAPEX) on HPC will pay off in expense savings

Enterprise Class IT

MDP challenges to Enterprise Class IT Infrastructure



- Size of files drives distributed processing
- Simultaneous read from input file by all “workers” bogs down file server
- Large CPU Count Fractures overtax NFS server CPU



- Maximum Read Bandwidth is 90 MB/sec

SYNOPSYS®
Predictable Success

Enterprise Class IT

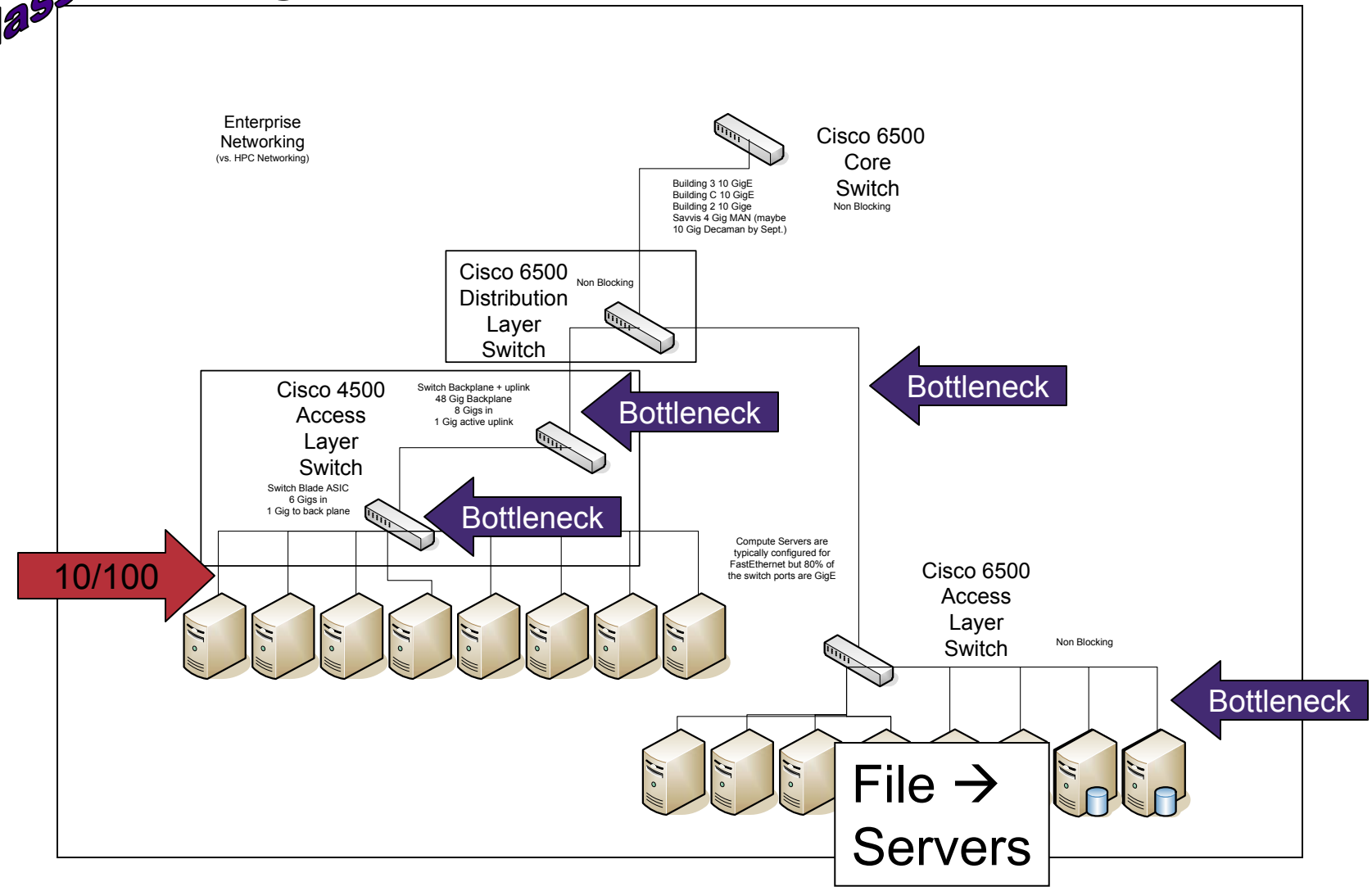
Blocking Factor

Some of the bottlenecks found in a typical enterprise data center.

- Computer Servers connected by Fast, rather than Gig Ethernet
- Access Layer switches with internal blocking factors
- Blocking factors created by multiple Ethernet switches, where the uplink bandwidth is less than the sum of the aggregate port bandwidth
- Network file servers connected with less bandwidth than the aggregate sum of the compute server bandwidths

Enterprise Class IT

Typical Enterprise Network



SYNOPSYS®
Predictable Success

Enterprise Class IT

Blocking Factor

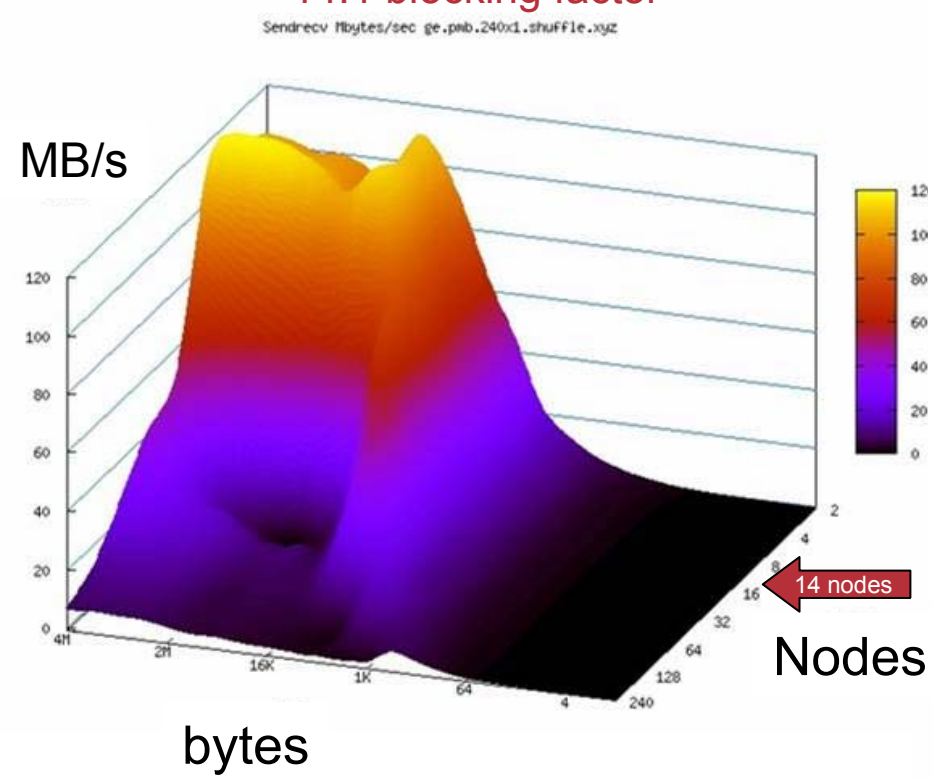
- All of these bottlenecks and blocking factors result in DP workers that cannot talk simultaneously, and at wire speed, to the storage server.
- This can be graphically illustrated with parallel IO tests such as the Pallas Parallel benchmark.

Enterprise Class IT

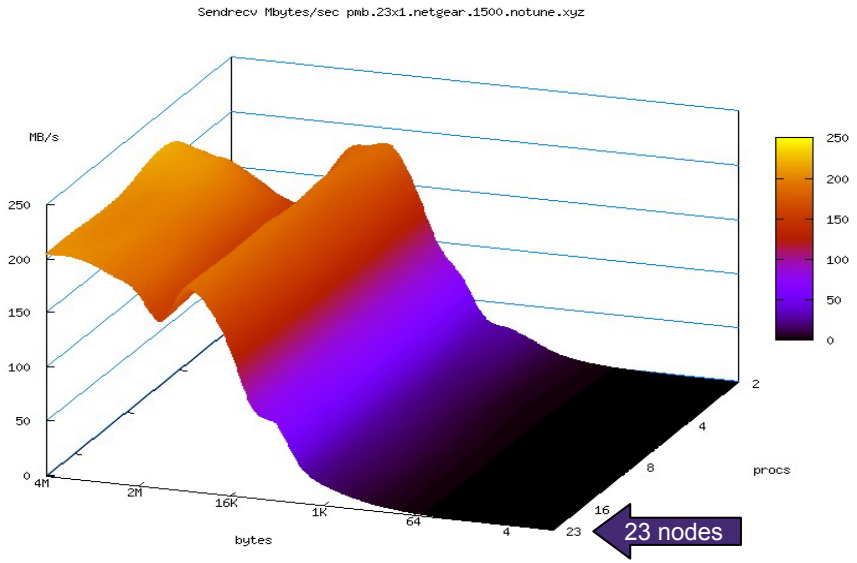
Blocking vs. Non-Blocking GigE

IBM Blade Centers

14:1 blocking factor

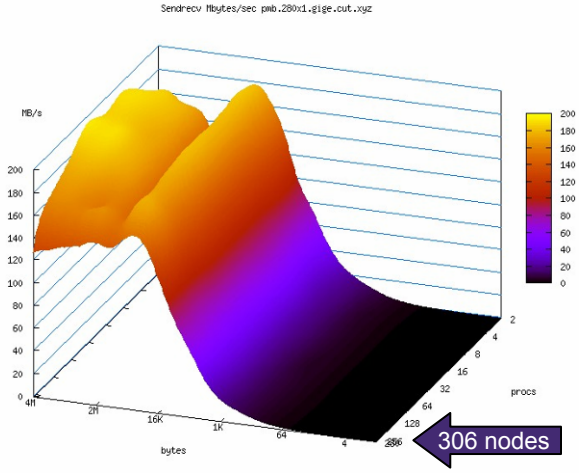


23 Nodes on Netgear JGS524 (24 Port GigE)



306 Nodes on Cisco Catalyst 6500

(Supervisor 720 & 6748 GigE blade)



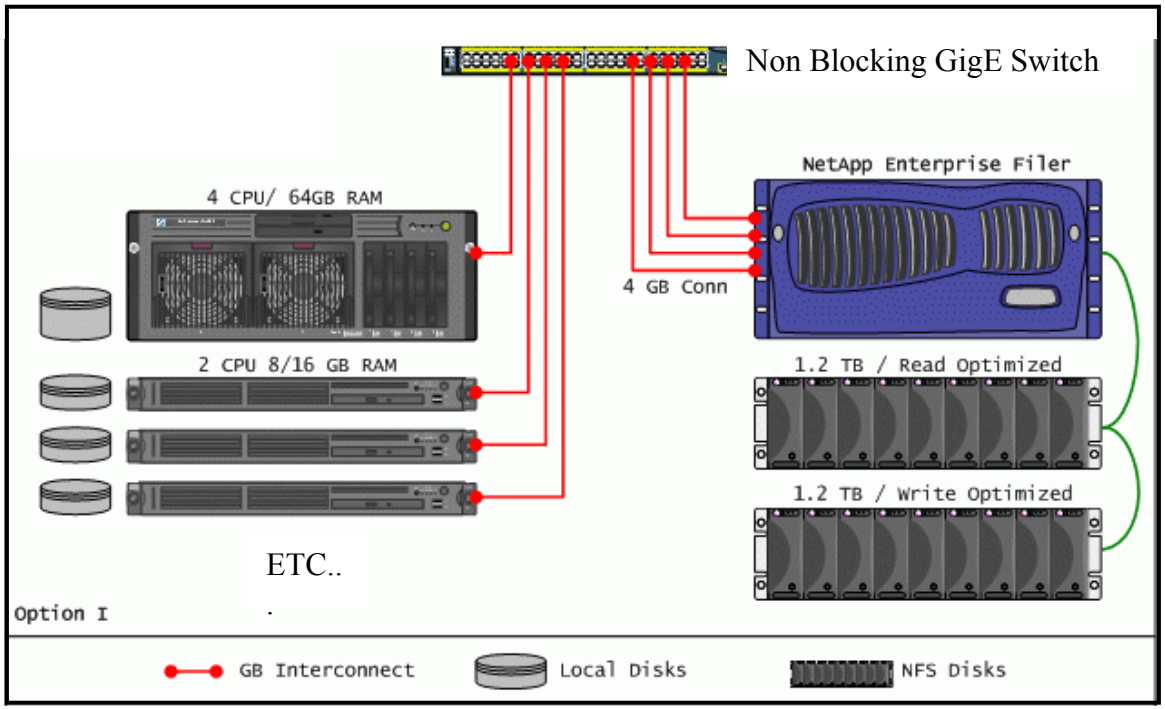
SYNOPSYS®
Predictable Success

Enterprise Class IT

GigE + NFS recommendations

For “Enterprise” class systems, the following recommendations can be used in order to mitigate the above performance issues (see also CATSTM DP User Guide Chapter 2).

- Use a dedicated Enterprise class NFS appliance, such as Network Appliance FAS 3020 or higher
- Make as many network connections to the NFS appliance as it allows
- Use a non-blocking, single GigE switch to connect DP Masters, workers, and storage server together.
- Call this 1X.



SYNOPSYS® Predictable Success

Going Beyond Enterprise Class

Interconnect

- There are three factors that affect performance over a given network interconnect.
- Bandwidth
- Latency
- CPU overhead

Interconnect	FastE	GigE	10GigE	Myrinet	Infiniband 4X SDR
Bandwidth	100 Mb	1 Gb	10 Gb	2 Gb	10 Gb
Latency	1.2ms	60us	10us	8us	4us
CPU overhead	80%	80%	80%	6%	3%

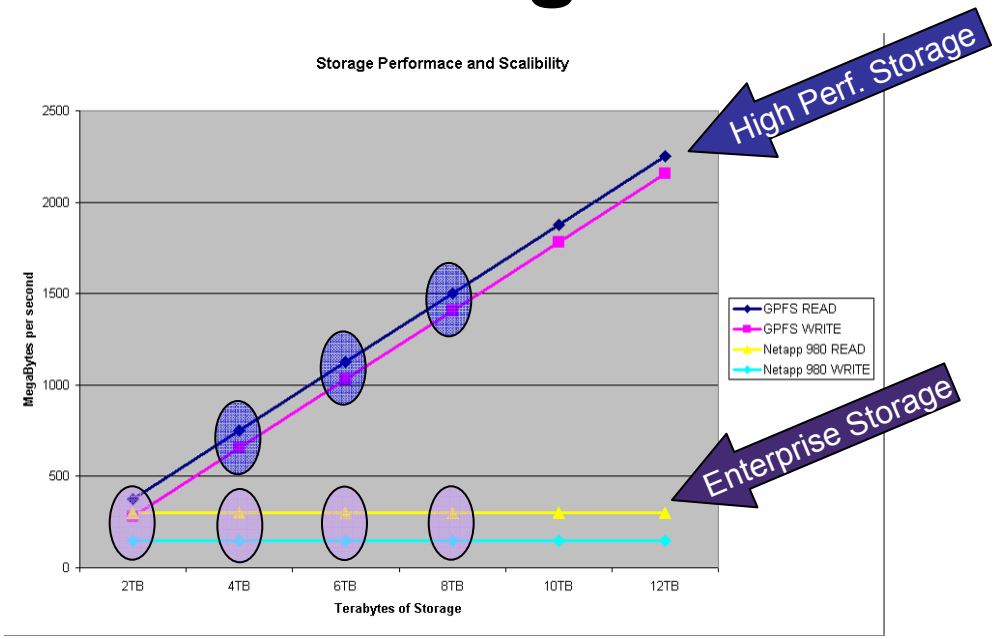
Of these interconnects, Infiniband has the best or equal performance in all three categories

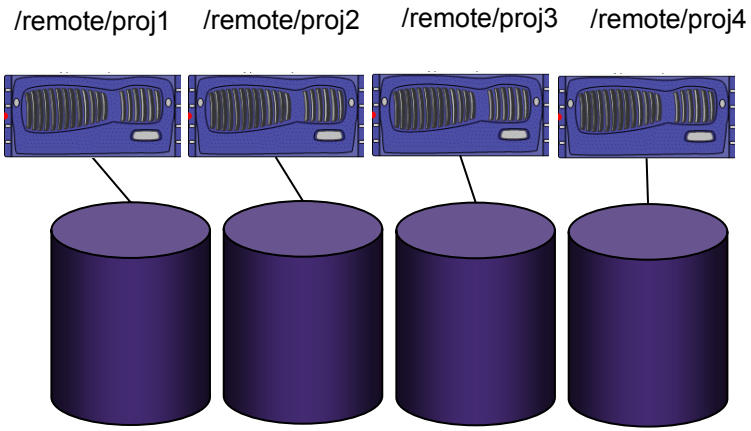
Economic Considerations

- While the “per port” cost of Infiniband Host adapters, cables, and switches may be higher than some of the other interconnects, the resulting performance offsets these costs by achieving equivalent turn around times with fewer compute nodes.

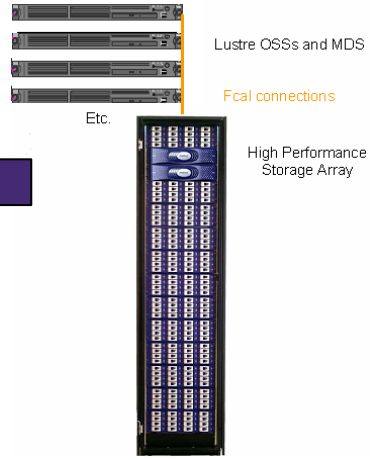
High Performance Storage

- Enterprise Storage
  - Aggregate Performance Increases as you add servers





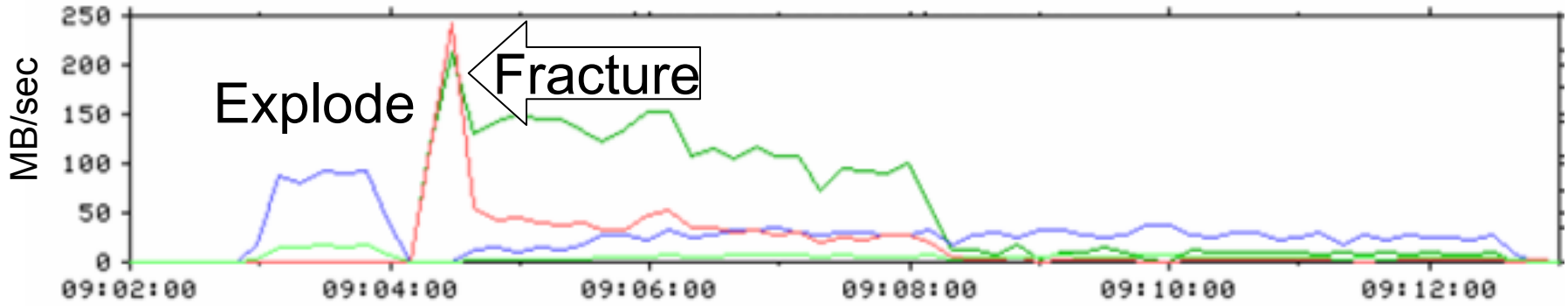
/proj



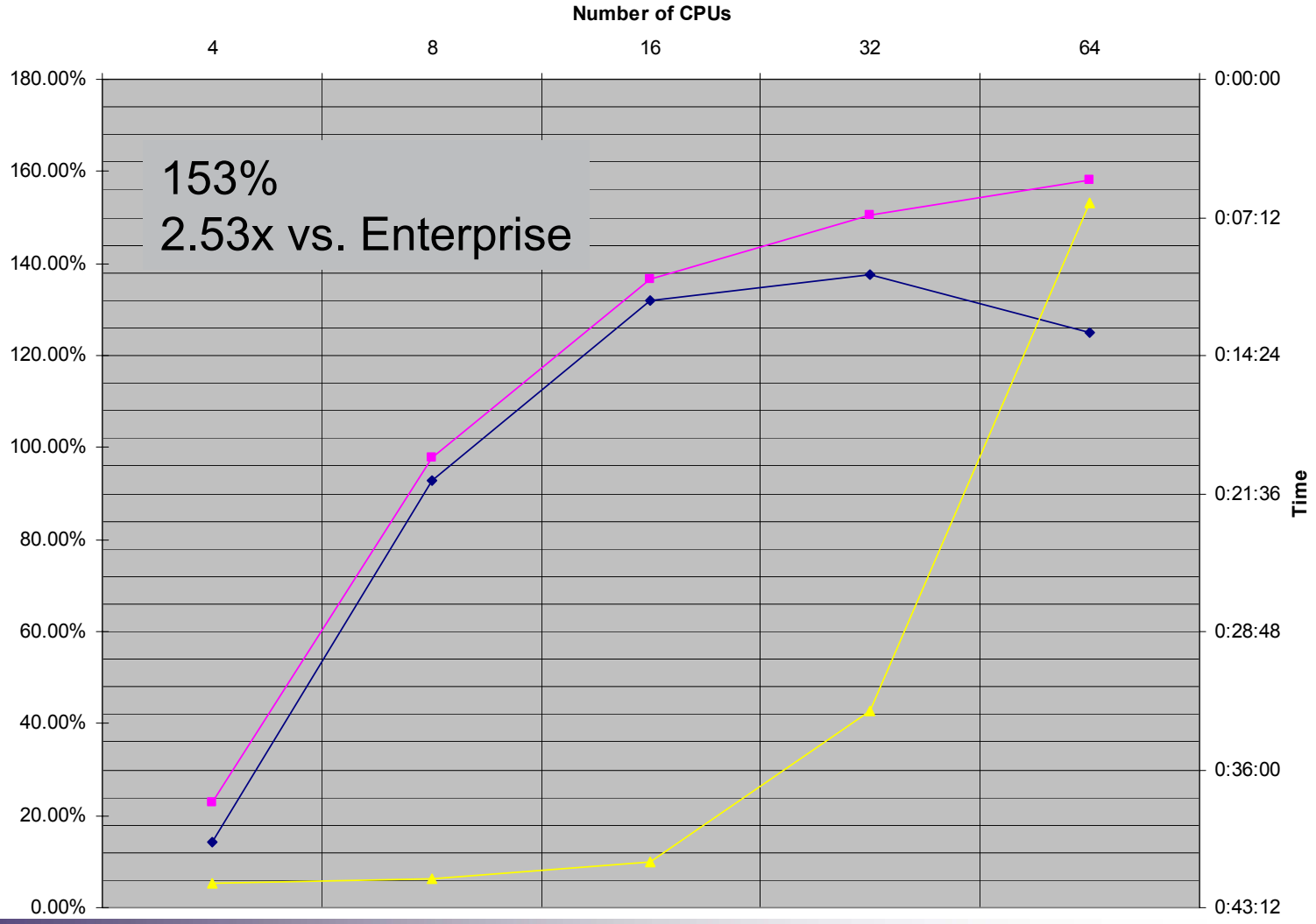
SYNOPSYS®
Predictable Success

Lustre File System

- Lustre is a parallel file system
- Lustre “speaks” Infiniband verbs
  - CATS™ Application can talk directly to storage without TCP overhead
  - No context switch needed
- Lustre used on many of the top super computers
- Tests showed 250 MB/s performance for single server, vs. 90MB/s NFS



Prototype Lustre Results



SYNOPSYS®
Predictable Success

The Final Solution

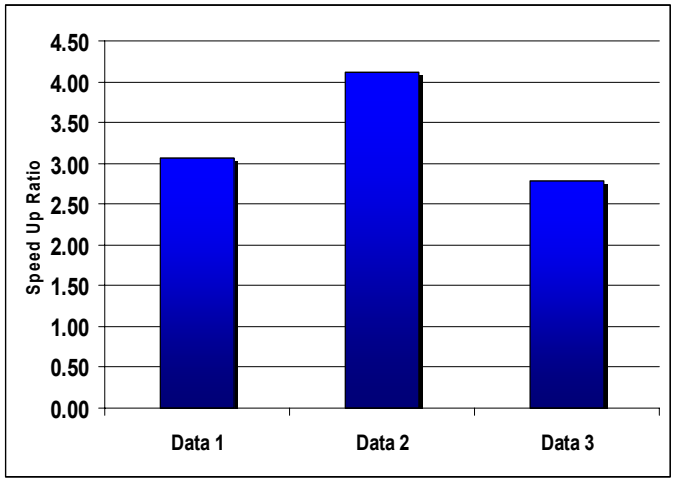
- For the final solution the storage system was designed to deliver the same amount of per CATSTM worker process storage performance as seen in the prototype cluster, but scaled to 256 worker processes.
- 17x the storage performance of the GigE + NFS cluster
- 6x the performance of the prototype Infiniband + Lustre cluster
- 8 Lustre Object Storage Servers vs. 3
- 160 FCAL spindles 16 TB usable vs. 72 spindles 8TB
- Fault tolerant high performance storage array with 8x2 RAID stripe (can survive and perform through multiple disk failures) vs. software raid
- 264 port Infiniband switch with 6 GigE uplinks vs. 24 IB ports

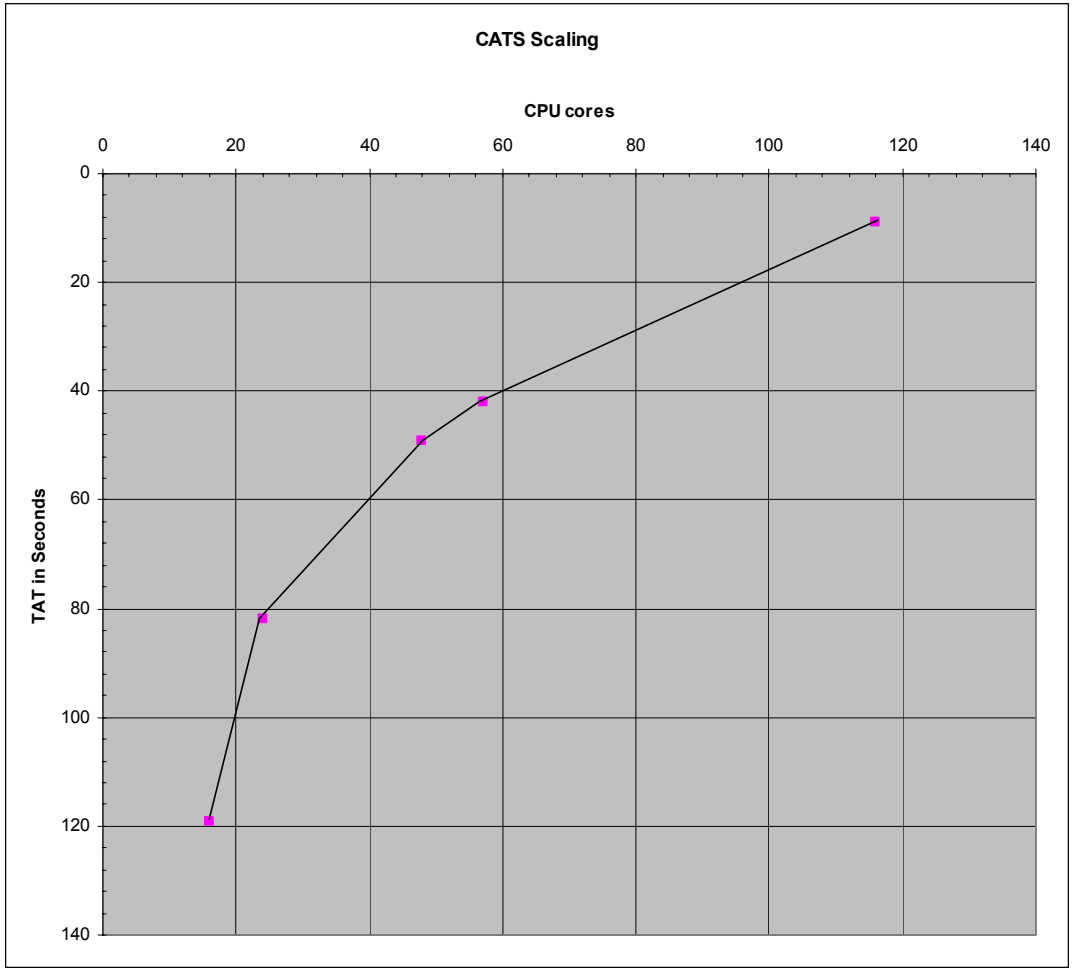
Demonstrated Scaling past 100 CPUs

Synopsys CATS™

New HPC Cluster Solution Delivers 4x Speedup

test run results on three data sets, comparing Lustre + Infiniband to GigE + NFS. Twenty CPU CATSTM parallel distributed processing was used on both clusters. The Data sets consisted of 300GB GDS files.
Speedups between 2.75x and > 4X are shown. The speedup factors will increase with larger numbers of CPUs





SYNOPSYS®
Predictable Success

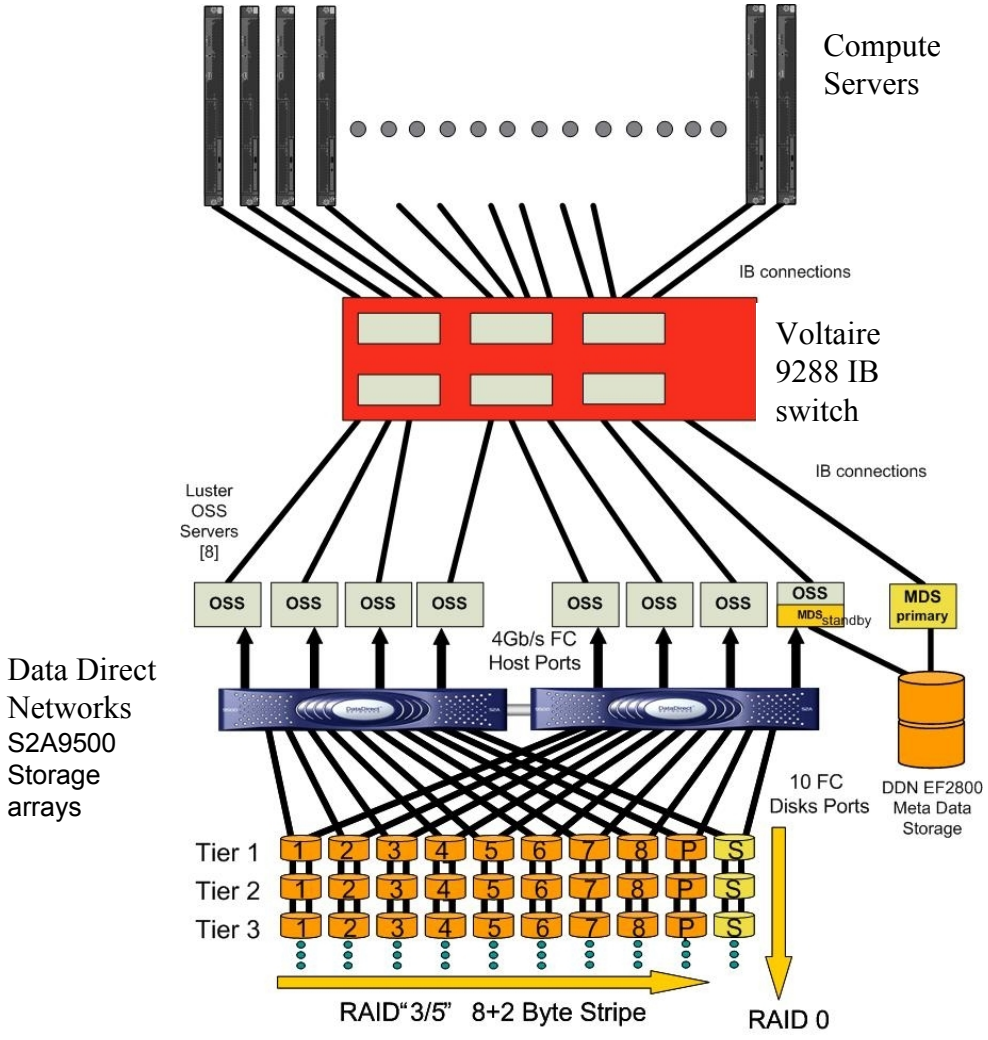
HPC Backend Storage

Super Computer Class Storage Performance – Enterprise Price

- In order to ensure the storage performance would scale, a “best in class” high performance storage array was chosen, based on it being deployed in 7 of the TOP 10 super computers (at that time), many of which use the Lustre file system.
- Interestingly, the per terabyte cost of this HPC storage was less than that of the normal Enterprise NFS solution.
- The idea is create a backend storage system that can saturate the interconnect

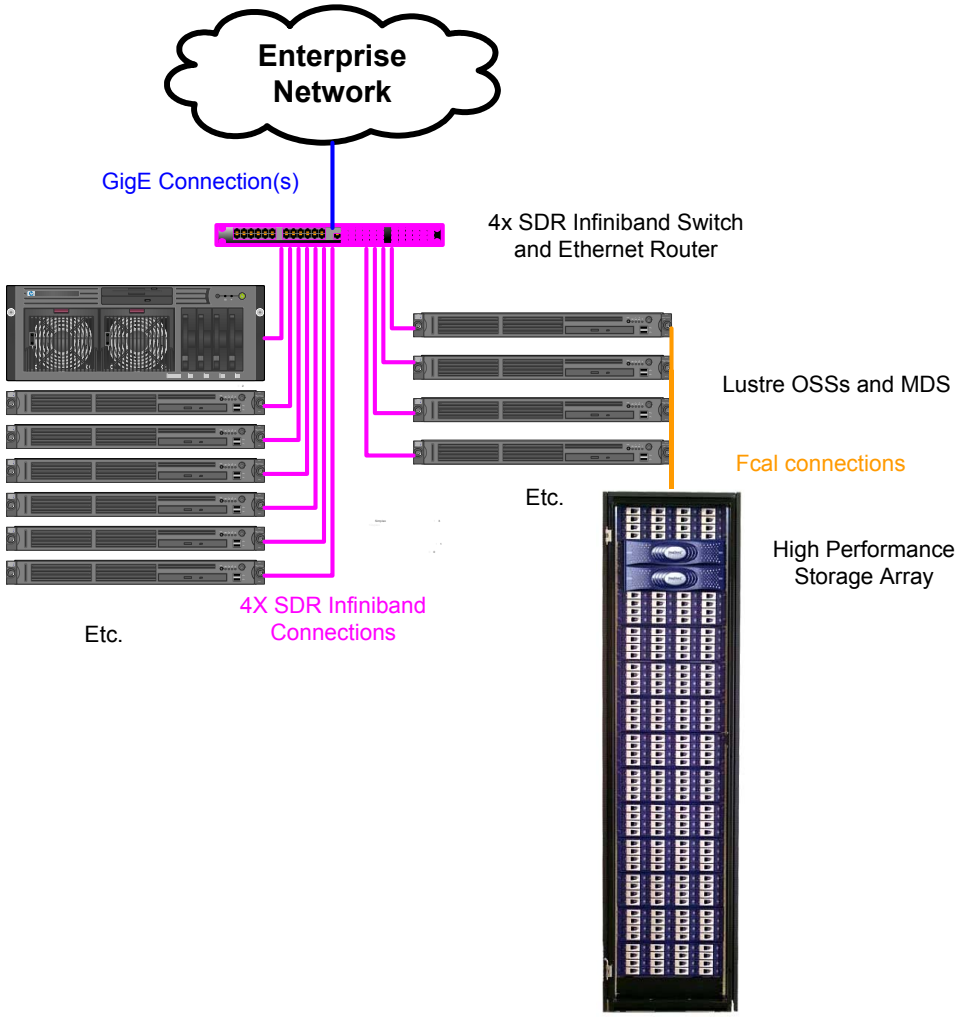
Schematic View

For the final solution the storage system was designed to deliver the same amount of per CATS™ worker process storage performance as seen in the prototype cluster, but scaled to 256 worker processes.



Architectural View

- 17x the storage performance of the GigE + NFS cluster
- 6x the performance of the prototype Infiniband + Lustre cluster
- 8 Lustre Object Storage Servers vs. 3
- 160 FCAL spindles 16 TB usable vs. 72 spindles 8TB
- Fault tolerant high performance storage array with 8x2 RAID stripe (can survive and perform through multiple disk failures) vs. software raid
- 264 port Infiniband switch with 6 GigE uplinks vs. 24 IB ports



Best Practices for Maximum CATS™ performance

- Use a Linux based 64bit X86 cluster
- Use a parallel file system such as Lustre
- Use a high speed low latency non-blocking interconnect such as 4x Infiniband
- Use a high performance storage back end such as Data Direct Networks with lots of FCAL spindles
- Scale storage backend, storage servers, and interconnect so that the storage system can deliver a minimum of 15 - 25MB/sec (B for Bytes) per compute node (e.g. for 200 compute nodes storage and interconnect should be able to deliver 3 – 6 GB/sec)
  - Note: GigE bandwidth max = 125 MB/sec
  - 10 GigE/Infiniband 4x max = 1.25 GB/sec
  - However Infiniband has lower latency (than 10GigE) and other advantages such as the ability of Lustre to use native Infiniband protocols, bypassing the Linux Kernel and TCP/IP stack.

Getting the message out

- Press releases and partnerships with vendors
- “Assisting” vendors in supporting interconnect options
- Speaking at conferences
- Communicating with our customers

Predictable Success