Technology and Architectures for Future Large-Scale Computing Systems

Rick Stevens
Argonne National Laboratory
The University of Chicago

Tools have changed rapidly: power

These were our supercomputers in the 1970's and 1980's

1986:
X-MP/48 ~220 Mflop sustained
120-150kW (depending on model)
$40M for computer+disks (FY09$)

Today:

NNSA:
Roadrunner at 1.105 PF (LINPACK)
LANL; 2.5 MW

SC/ASCR:
Jaguar at 1.059 PF (LINPACK)
ORNL; 6.9 MW

Factor 5×10^6 in speed
Factor of 18 in power

U.S. DEPARTMENT OF ENERGY

Argonne's IBM Blue Gene/P – 556 TFs

Cray XT5 at ORNL – 1.6 Pflop/s

Jaguar	Total	XT5	XT4
Peak Performance	1,645	1,382	263
AMD Opteron Cores	**181,504**	**150,176**	**31,328**
System Memory (TB)	**362**	**300**	**62**
Disk Bandwidth (GB/s)	**284**	**240**	**44**
Disk Space (TB)	**10,750**	**10,000**	**750**
Interconnect Bandwidth (TB/s)	**532**	**374**	**157**

The systems will be combined after acceptance of the new XT5 upgrade. Each system will be linked to the file system through 4x-DDR Infiniband

Traditional Sources of Performance Improvement are Flat-Lining (2004)

- New Constraints
  - 15 years of *exponential* clock rate growth has ended
- Moore's Law reinterpreted:
  - How do we use all of those transistors to keep performance increasing at historical rates?
  - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!

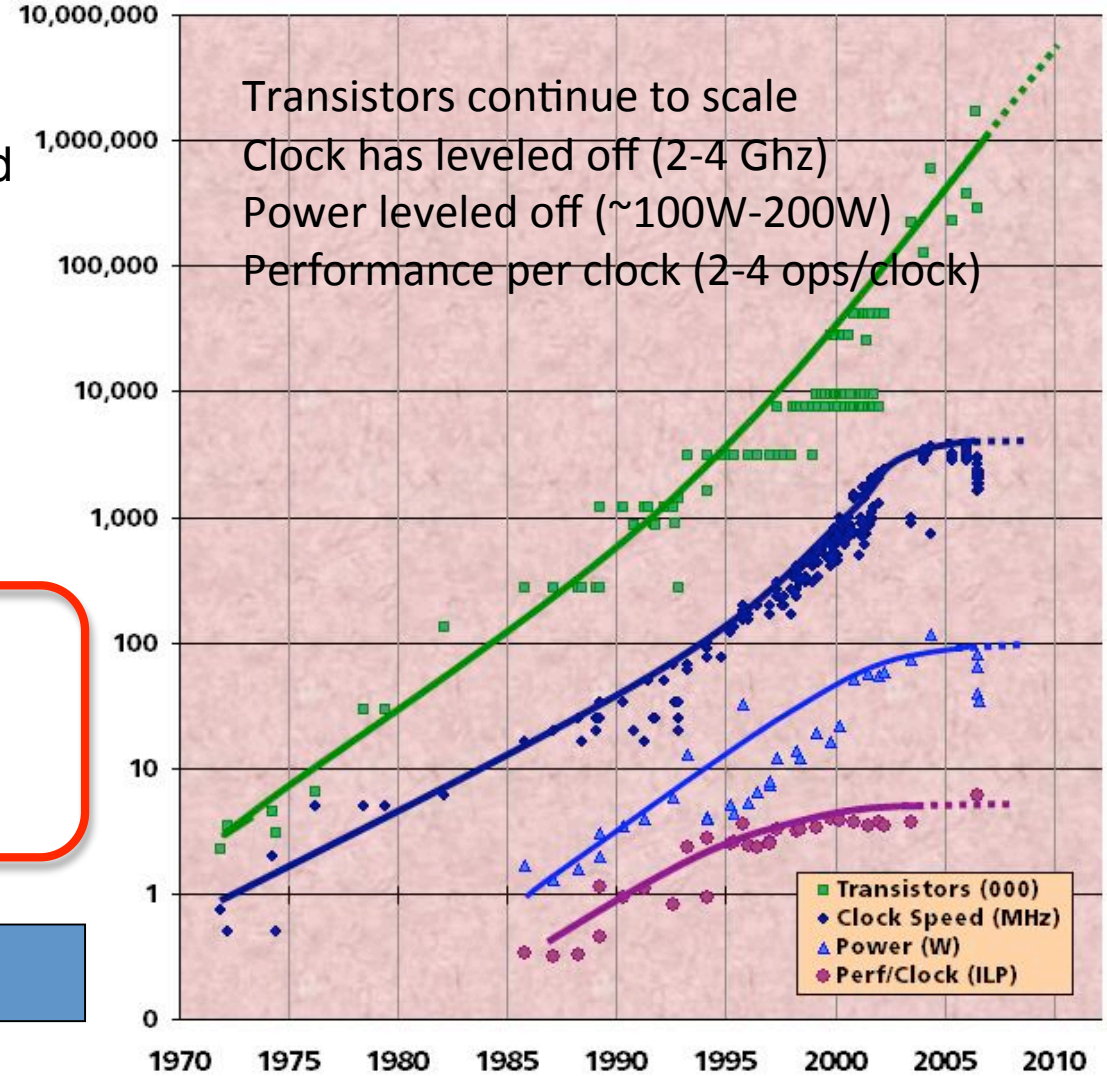
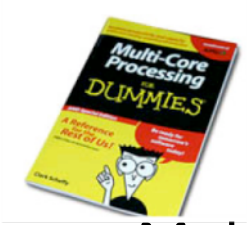
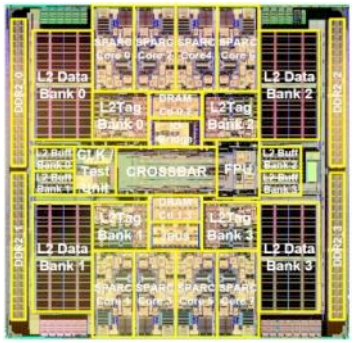


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith



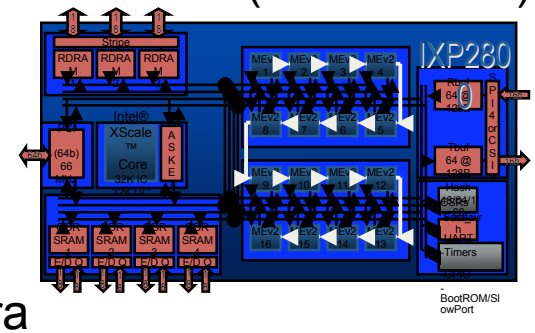
Multicore comes in a wide variety

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

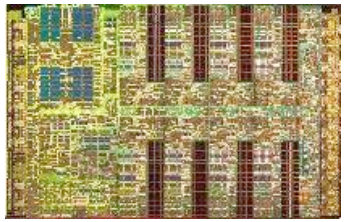


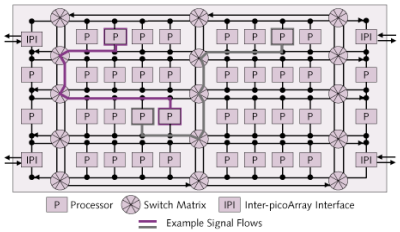
Sun Niagara
8 GPP cores (32 threads)

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



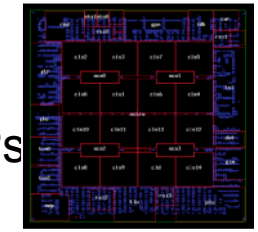
IBM Cell
1 GPP (2 threads)
8 ASPs

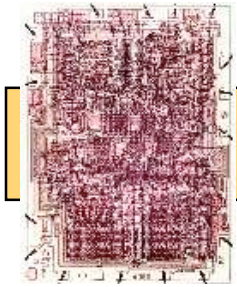




Picochip DSP
1 GPP core
248 ASPs

Cisco CRS-1
188 Tensilica GPPs



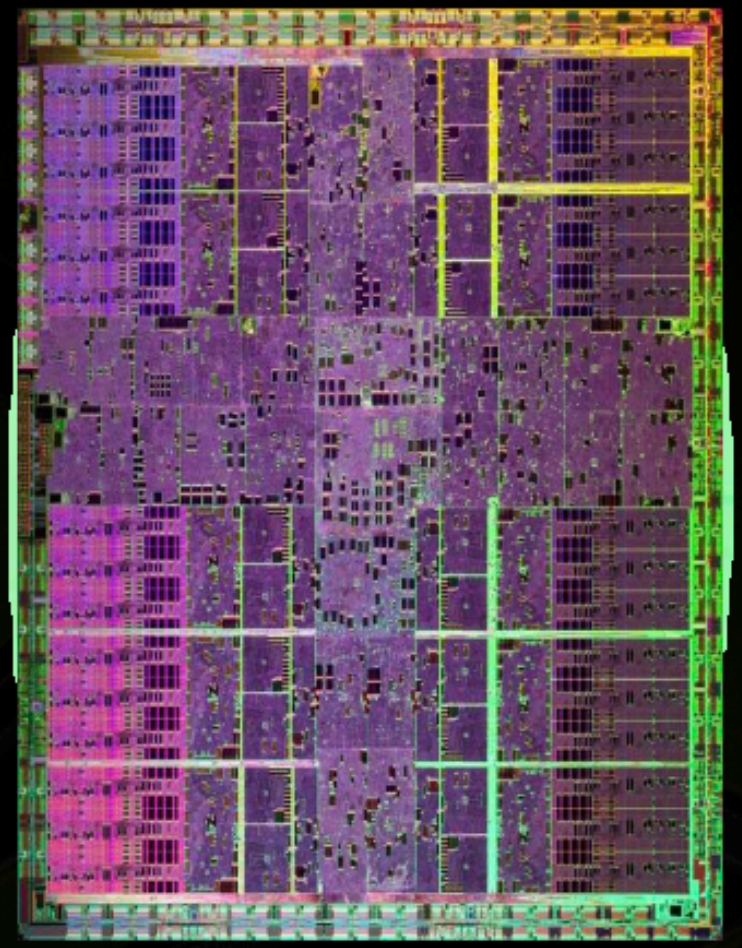


Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm^2 chip

"The Processor is the new Transistor"

[Rowen]

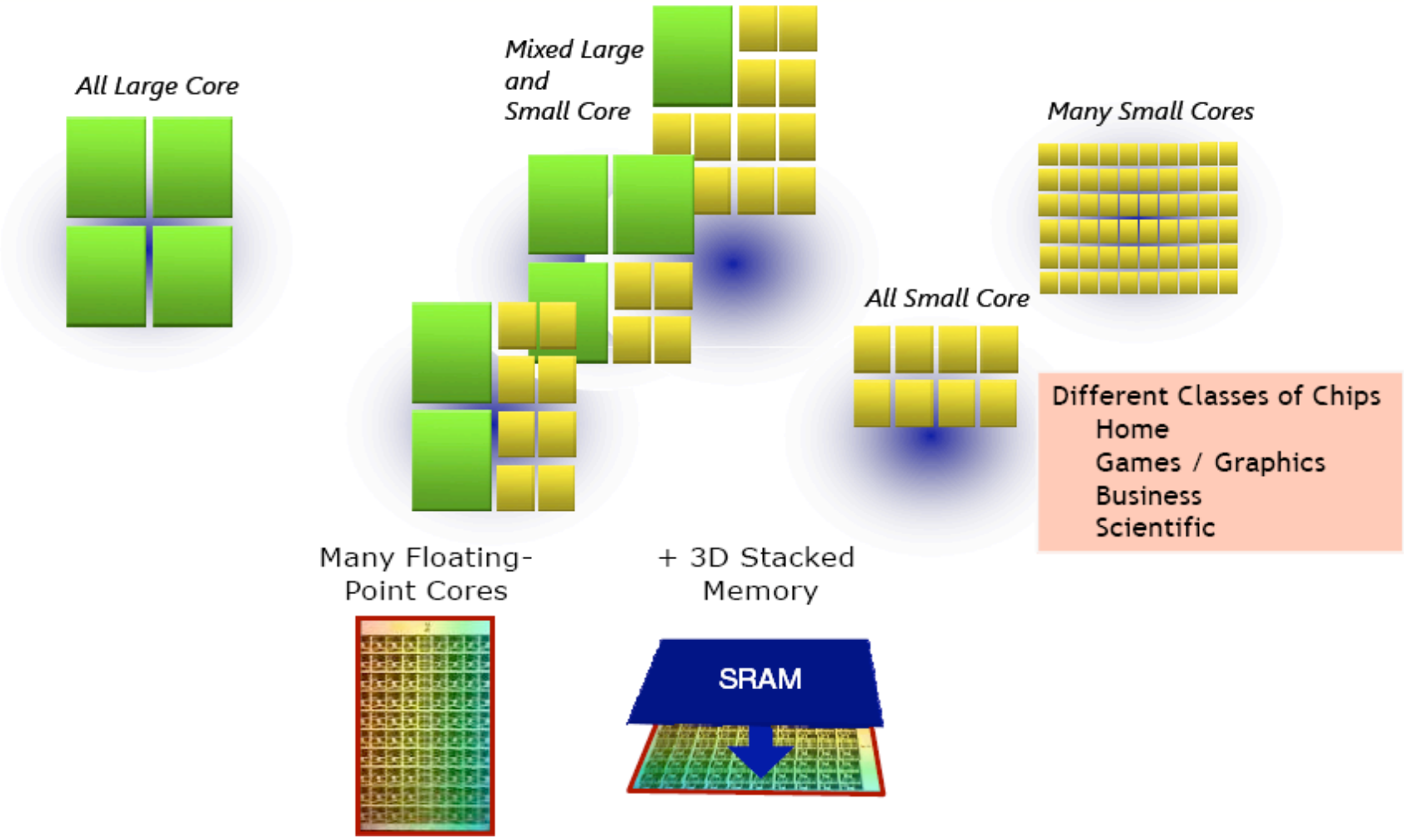
240 Cores Today, 1000s Soon



NVIDIA GT200 GPU

240 32b/64b FP Cores

What's Next?



Source: Jack Dongarra, ISC 2008

E3 Advanced Architectures – Findings (2007)

- Exascale systems are likely feasible by 2017±2
- 10-100 Million processing elements (mini-cores) with chips as dense as 1,000 cores per socket, clock rates will grow slowly
- 3D chip packaging likely
- Large-scale optics based interconnects
- 10-100 PB of aggregate memory
- > 10,000's of I/O channels to 10-100 Exabytes of secondary storage, disk bandwidth to storage ratios not optimal for HPC use
- Hardware and software (OS) based fault management
- Simulation and multiple point designs will be required to advance our understanding of the design space
- Achievable performance per watt will likely be the primary metric of progress

Top Technical Challenges

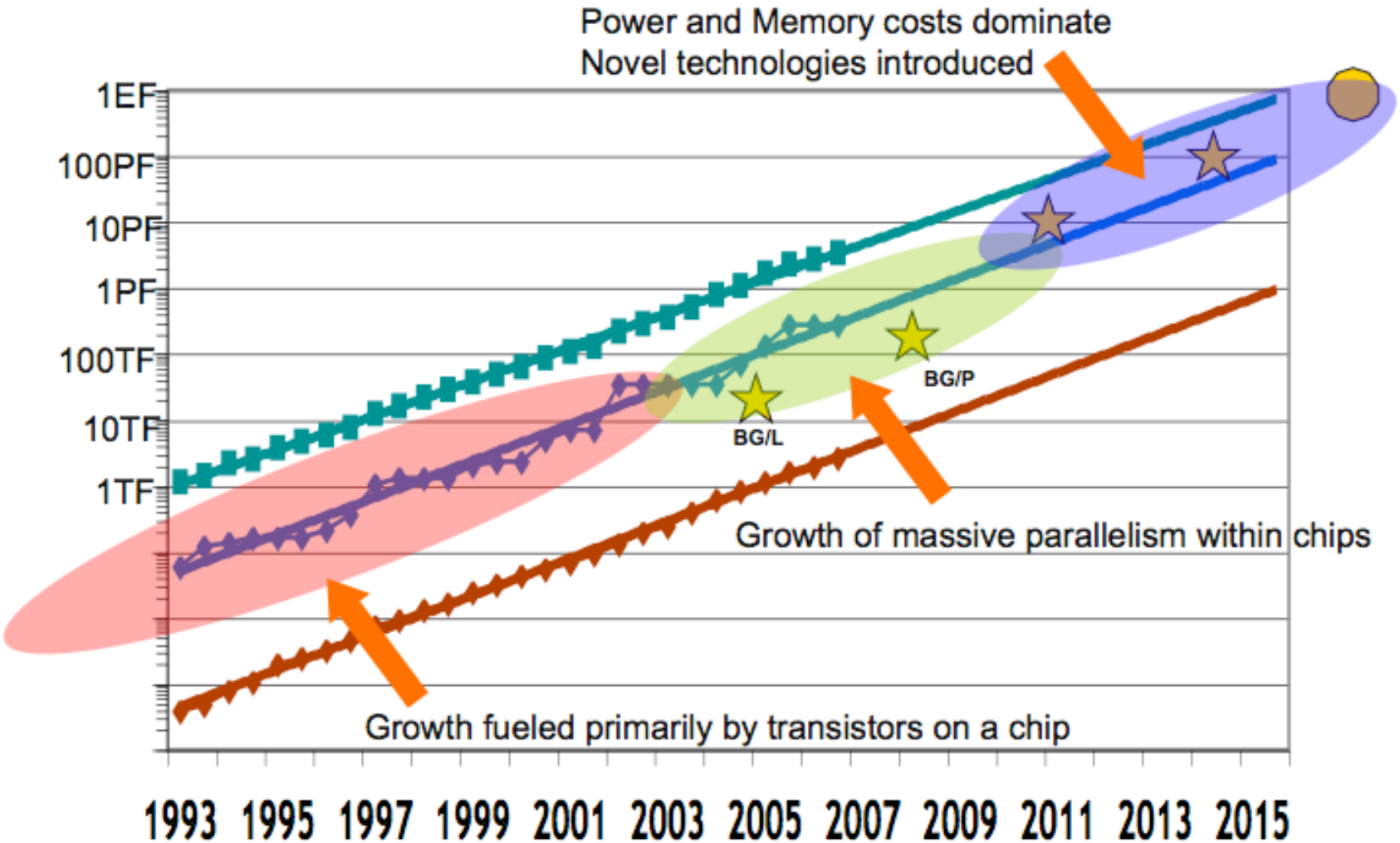
- Power Consumption
  - Proc/mem, I/O, optical, memory, delivery
- Chip-to-Chip Interface Scaling
  - pin/wire count ⇒ 3D packaging
- Package-to-Package Interfaces
  - Scalable interconnects
- Cost Pressure in Optics and Memory
- Fault Tolerance
  - FIT rates and fault management concepts
    - E.g. In place spares, chip kill, etc.
  - Reliability of irregular logic, design practice

Exascale systems built from today's technology would consume over a Gigawatt

Exascale systems built from today's technology would cost over a Gigabuck

Exascale systems built from today's technology would have a MTBF of 10 minutes

Looking out to Exascale
Concurrency will be Doubling every 18 months



Systems Scaling Projections

Begin Full System Delivery (Yr)	2004	2007	2012	2015	2019
Design Parameters	**BG/L**	**BG/P**	**25PF**	**300PF**	**1200PF**
Cores / Node	2	4	8-24	32-64-128	96-128-500
Clock Speed (GHz)	0.7	0.85	1.6-4.1	2.3-4.8	2.8-6.0
Flops / Clock / Core	4	4	8-32	8-32	16-64
Nodes / Rack	1024	1024	100-1024	256-1024	256-1024
Racks / Full System Config	64	72	128-350	128-400	256-400
MB RAM/core	256	512	1024-4096	1024-4096	1024-4096
Total Power	2.5MW	4.8MW	8MW-20MW	20MW-50MW	30MW-80MW
Flops / Node (GF)	5.6	14	128-640	640-2000	2000-6000
Flops / Rack (TF)	5.7	14	200-400	400-1200	1600-4800
LB Concurrency	5.E+05	1.E+06	1M-2M	10M-100M	400M-1000M
Full System					
Total Cores (Millions)	0.13	0.3	.3M-1.2M	1M-10M	4M-200M
Total RAM (TB)	33.6	151	2,000-4,400	3,000-10,000	5,000-25,000
Total Racks	64	72	128-350	128-400	256-400
Peak Flops System (PF)	**0.37**	**1**	**25**	**300**	**1200**

ITRS Roadmap for Logic Devices

	Units	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Feature Size	nm	90	78	68	59	52	45	40	36	32	28	25	22	20	18	16	14
Logic Area	relative	1.00	0.80	0.63	0.51	0.39	0.32	0.25	0.20	0.16	0.12	0.10	0.08	0.06	0.05	0.04	0.03
SRAM Area	relative	1.00	0.78	0.61	0.48	0.38	0.29	0.23	0.18	0.14	0.11	0.09	0.07	0.06	0.04	0.03	0.03
50/50 Area	relative	1.00	0.79	0.62	0.49	0.38	0.30	0.24	0.19	0.15	0.12	0.09	0.07	0.06	0.05	0.04	0.03
High Performance Devices																	
Delay	ps	0.87	0.74	0.64	0.54	0.51	0.40	0.34	0.29	0.25	0.21	0.18	0.15	0.13	0.11	0.10	0.08
Average Device Capacitance	relative	1.00	0.87	0.76	0.66	0.58	0.50	0.44	0.40	0.36	0.31	0.28	0.24	0.22	0.20	0.18	0.16
Circuit speedup: 1/delay	relative	1.00	1.18	1.36	1.61	1.71	2.18	2.56	3.00	3.48	4.14	4.83	5.80	6.69	7.91	8.70	10.88
ITRS Max Clock	relative	1.00	1.30	1.78	2.11	2.38	2.90	3.39	3.86	4.42	5.45	6.42	7.63	8.75	10.22	12.00	14.05
Vdd	volts	1.10	1.10	1.10	1.00	1.00	1.00	1.00	0.90	0.90	0.90	0.80	0.80	0.70	0.70	0.70	0.70
Vdd/Vt	ratio	5.64	6.55	6.67	6.10	4.22	6.62	6.85	6.08	5.39	5.49	4.82	4.10	3.50	3.48	3.41	3.37
Power Density @ Circuit Speedup	relative	1.00	1.29	1.65	1.77	2.13	2.95	3.95	4.19	5.52	7.41	7.54	10.19	10.17	13.91	17.12	23.47
Power Density @ Max Clock	relative	1.00	1.43	2.17	2.31	2.97	3.93	5.23	5.39	7.00	9.75	10.01	13.39	13.30	17.98	23.61	30.33
Energy/Operation	relative	1.000	0.867	0.756	0.542	0.478	0.413	0.367	0.268	0.238	0.208	0.147	0.129	0.090	0.081	0.072	0.063
Low Operating Power Devices																	
Delay	ps	1.52	1.33	1.17	1.03	0.90	0.79	0.79	0.61	0.53	0.47	0.41	0.36	0.32	0.28	0.24	0.21
Circuit speedup: 1/delay	relative	0.57	0.65	0.74	0.84	0.97	1.10	1.10	1.43	1.64	1.85	2.12	2.42	2.72	3.11	3.63	4.14
Vdd	volts	0.90	0.90	0.80	0.80	0.80	0.70	0.70	0.70	0.60	0.60	0.60	0.50	0.50	0.50	0.50	0.50
Vdd/Vt	ratio	3.13	2.97	2.81	2.95	2.90	3.10	3.00	3.03	2.33	2.40	2.39	2.10	2.09	2.07	2.06	2.03
Power Density @ Circuit Speedup	relative	0.38	0.48	0.48	0.59	0.77	0.73	0.83	1.21	1.16	1.47	1.86	1.66	2.11	2.79	3.64	4.56
Energy/Operation	relative	0.669	0.588	0.468	0.347	0.306	0.202	0.188	0.162	0.106	0.093	0.083	0.051	0.046	0.041	0.037	0.032

Note: units of "relative" represent values normalized to those of the 2005 high performance technology

Figure 6.1: ITRS roadmap logic device projections

Power Density 20x-30x than 2005

Voltage assumption is .5v - .7v

Darpa Exascale Study

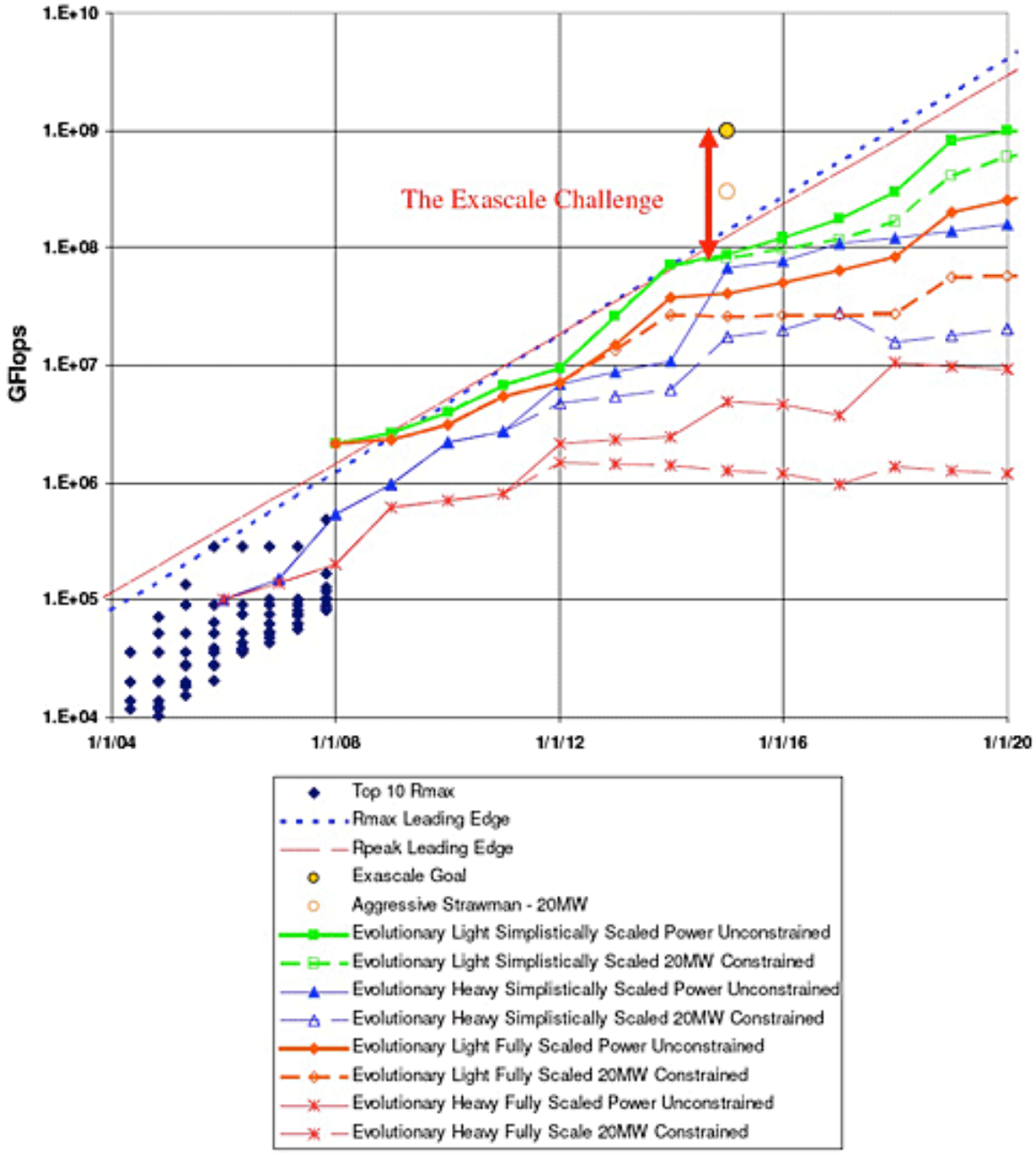
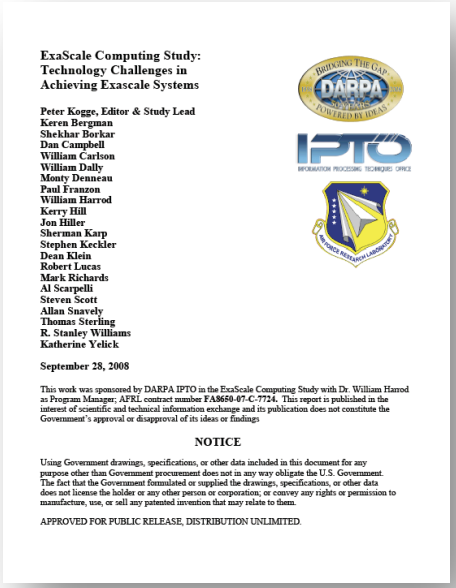


Figure 8.1: Exascale goals - Linpack.

Concluded that it will be a Major challenge to get to Sustained Exaops performance Levels by 2020



ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number FA8650-07-C-7724. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

Total System Concurrency

Largest machines now have combined system's level concurrency approaching 1M

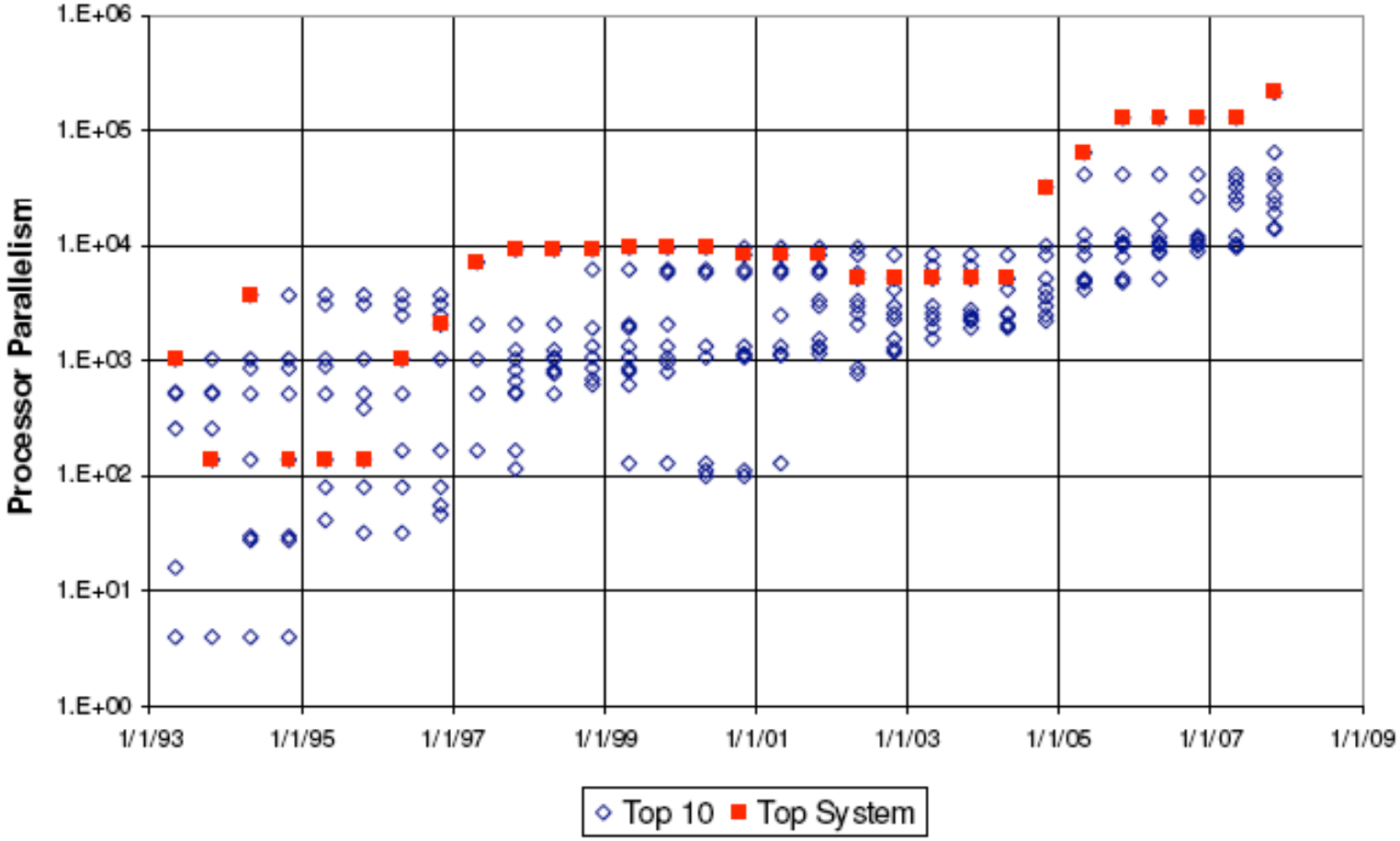


Figure 4.13: Processor parallelism in the Top 10 supercomputers.

Thread Level Concurrency

Each thread must do multiple operations per clock cycle

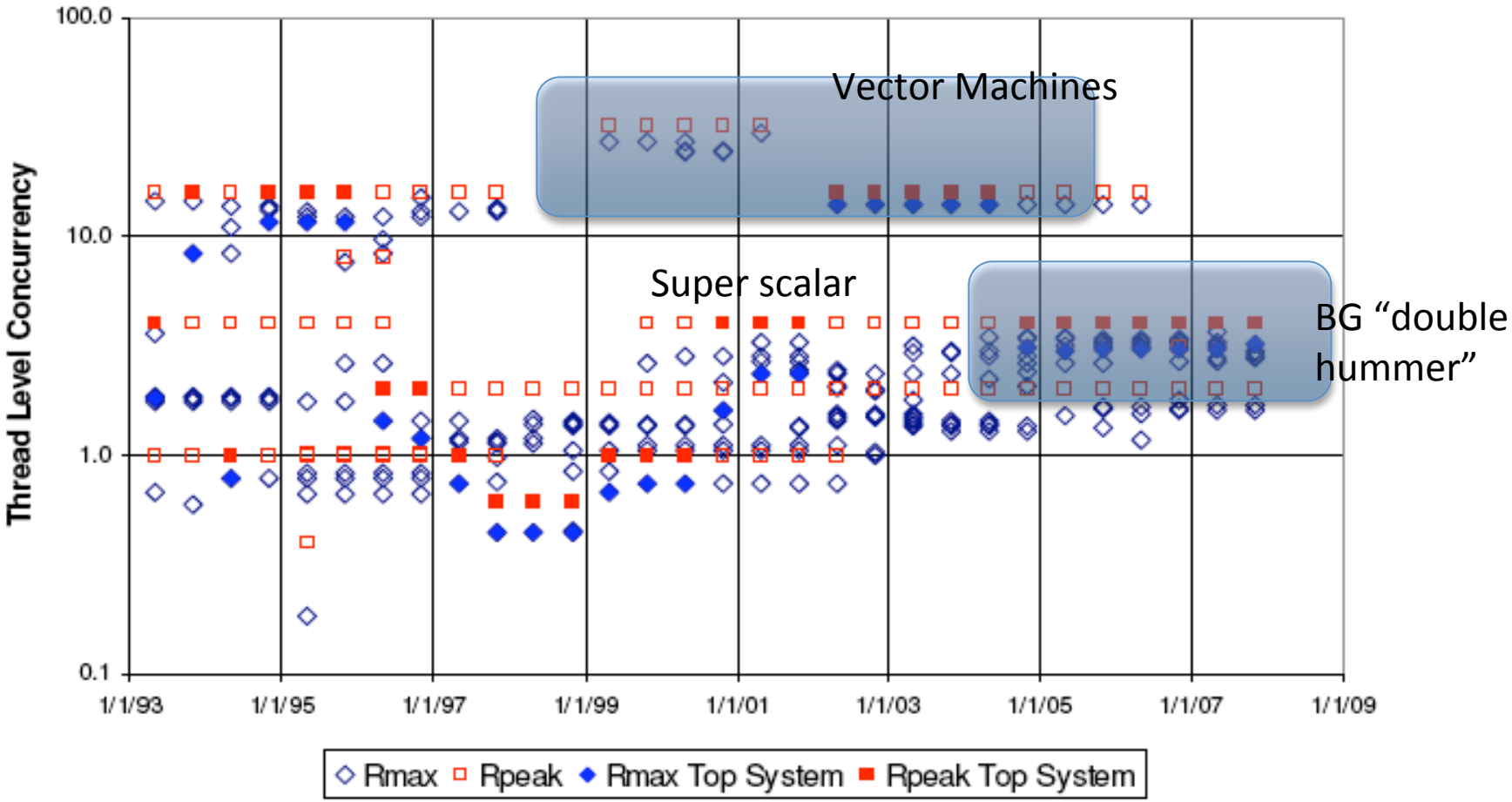


Figure 4.15: Thread level concurrency in the Top 10 supercomputers.

Parallelism and Locality Trends

Algorithms are pushing in the opposite direction as the hardware
e.g. advanced algorithms are becoming less "local" while
hardware is becoming more parallel and more local

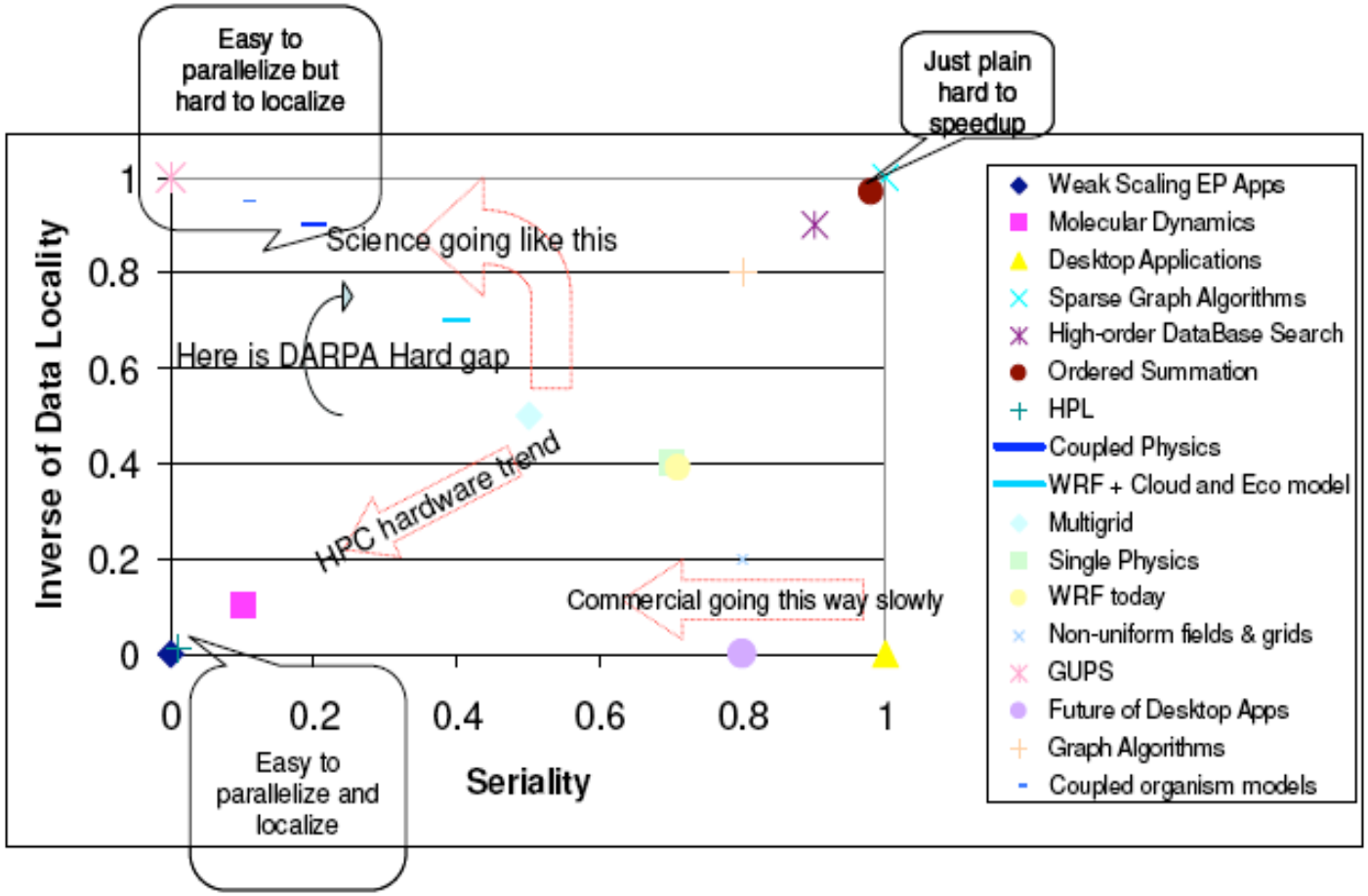


Figure 5.16: Future scaling trends

Power Constrained Clock Rate

Clock = Power_Density/ (Capacitance_per_device * Transistor_Density * V^2_{dd})

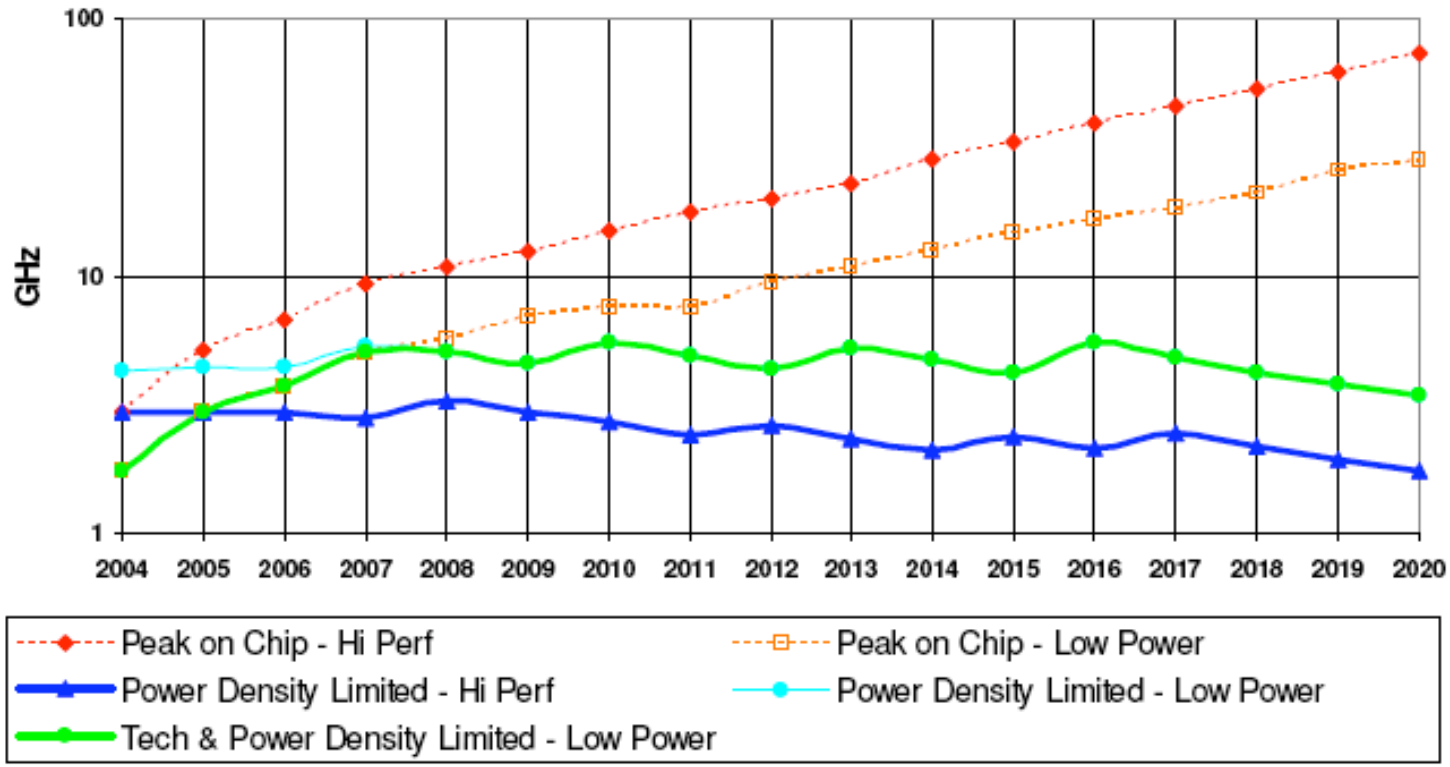


Figure 6.3: Power-constrained clock rate

Power density (i.e. cooling) will limit the clock rate as feature sizes shrink

Gflops per Watt must grow 1000x (0.1 ⇒ 100) to achieve Exascale

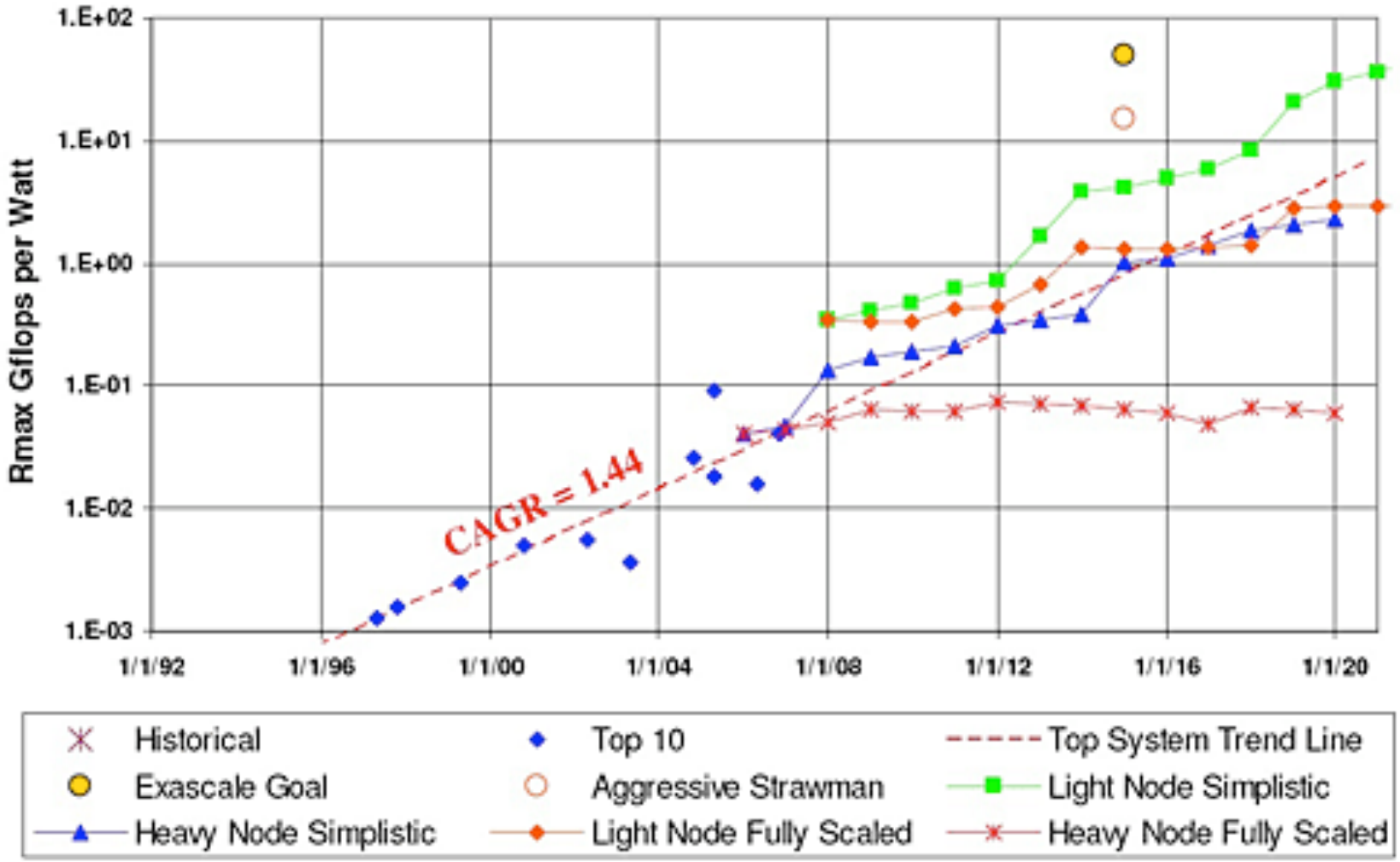
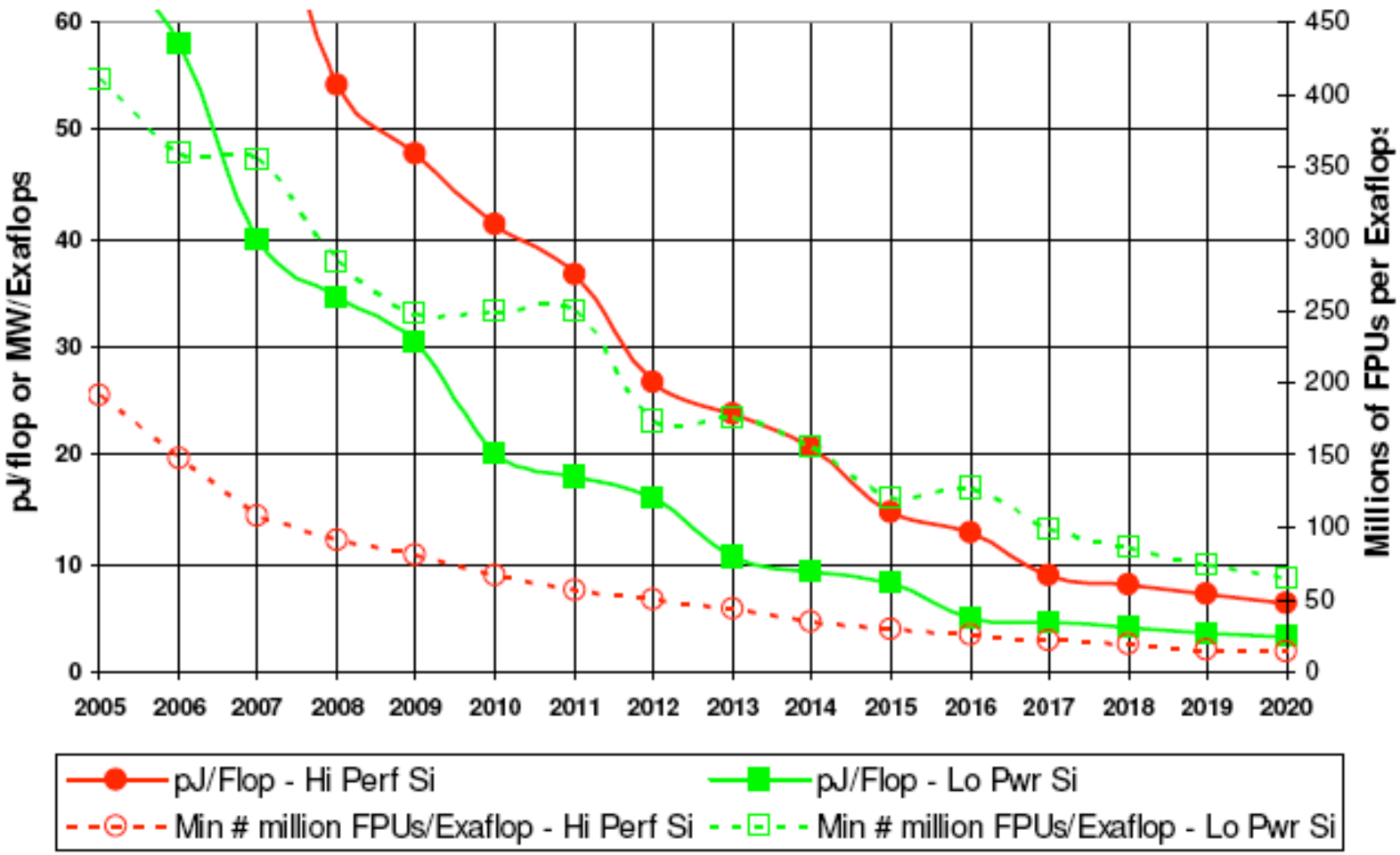


Figure 8.3: The power challenge for an Exaflops Linpack.

Power and FPU scaling needed to Reach Exa-operations/s (only flops)



Pico joule/FLOP

Figure 7.1: Projections to reach an Exaflop per second.

FPUs per Exaflop

Interconnect Technology Roadmap

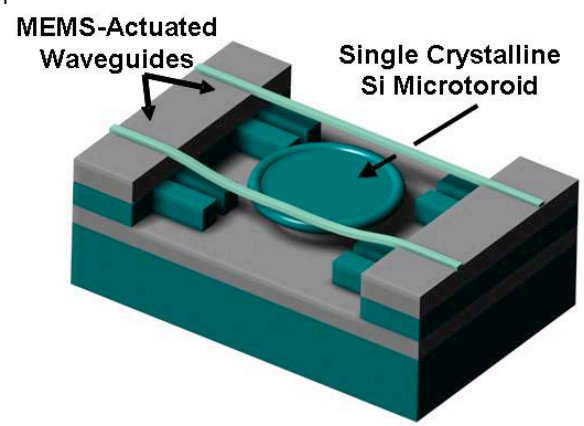
Technology	Density (wires/mm)	Power (pJ/bit)	Technology Readiness
Long-range on-chip copper	250	18 fJ/bit-mm	Demonstrated
Chip-to-chip copper	8	2 pJ/bit. Includes CDR	Demonstrated. Potential for scaling to 1 pJ/bit
Routed interconnect in 2015	n/a	2 pJ/bit router or non-blocking circuit switch 1 pJ/bit	roughly the same for packet
Optical State of Art (multi-mode)	10	9 pJ/bit. NOT including CDR	Demonstrated.
Optical (Single mode) in 2010	300	7.5 pJ/bit SOI waveguides PCB-embedded waveguide does not exist	Assumes lithographed
Optical (Single mode) in 2015	300	1.5 pJ/bit	At early research stag
Optical Routing		Add 0.1 pJ/bit (2010) for each switch	
Optical - temperature control			TEC cooler demonstrated
CNT bundles	1250	6 fJ/bit-mm	Undemonstrated

Table 6.8: Summary interconnect technology roadmap.

Long distance Interconnects Will be optical

Optical may replace Copper for on carrier Interconnects

Interconnect power will Grow to 10%-30% of total power

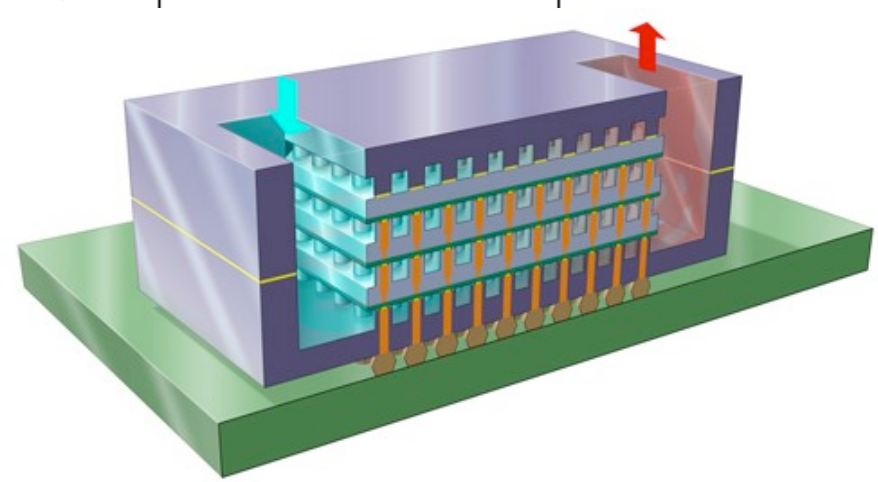


Heat Removal Approaches

Highly likely all future machines at scale will be liquid cooled

Approach	Thermal Performance	Comments
Copper Heat Spreader	Thermal conductivity = 400 W/(m.K)	
Diamond	Thermal conductivity = 1000 - 2000 W/(m.K)	Expensive
Heat Pipe	Effective conductivity = 1400 W/(m.K)	Very effective
Thermal Grease	Thermal conductivity = 0.7 - 3 W/(m.K)	
Thermal vias with 10% fill factor	Effective Conductivity = 17 W/(m.K)	
Thermal Electric Coolers	Limited to less than 10 W/cm^2 and Consumes Power	
Carbon Nanotubes	Excellent	

Table 6.9: Internal heat removal

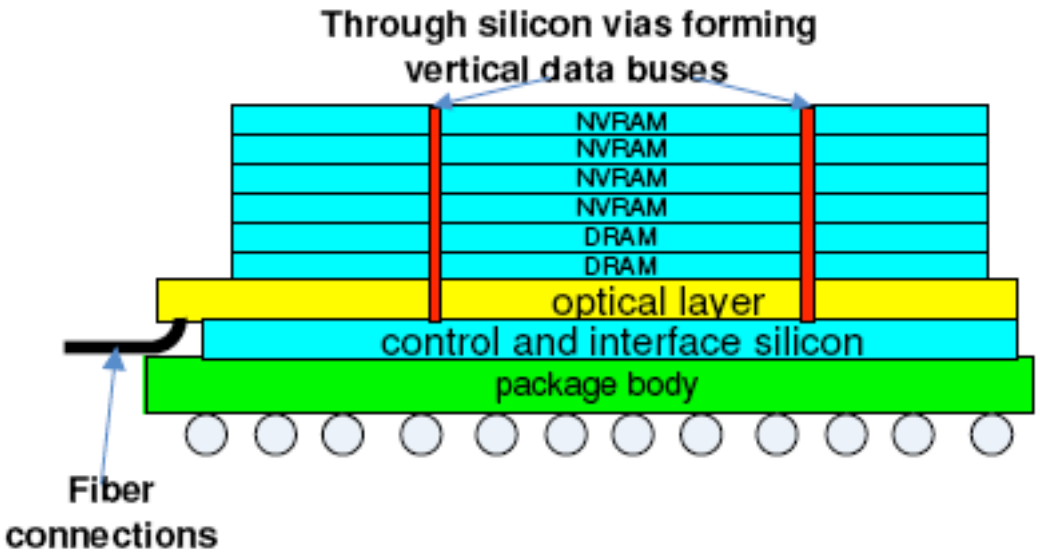


High-End Packaging Options

Approach	Wires/mm or sq.cm e.g. 2 wires/mm 1 mm	Bandwidth/mm Routable signal pairs per mm per layer * # layers * bit-rate per signal pair	Comments
Laminate (Ball Grid Array)	20 wires/mm/layer (2-4 signal layers) ~2,000 max total pin count	6 pairs/mm @ 30 Gbps = 180 – 720 Gbps/mm (1-4 signal layers) Package I/O: 500 pairs = 15 Tbps	1 mil line/trace presents practical limit. 1 mm BGA ball pitch
Silicon Carrier	50 wires/mm/layer (2 signal layers) Has to be packaged for I/O	12 pair/mm @ 30 Gbps = 360 – 720 Gbps/mm (1-2 signal layers)	2 signal layers is practical limit.
3DIC Stack	~10-40 wires/mm vertically around edge	Total: 100 – 200 pair @ 10 Gbps → 0.5 - 2 Tbps (assumes memory)	Limited interconnect performance
3D IC with Through Silicon Vias	In excess of 10,000 vias per sq.mm.	In excess of 100,000 Tbps/sq.cm. Really determined by floorplan issues	Chip stack limited to 4-8 chips, depending on thermal and other issues
Stacked Packages	1/mm on periphery	25 pairs total @ 10 Gbps → 250 Gbps	Not very applicable to high performance systems
Stacked Silicon Carriers	Vertical connections @ 20 um pitch → 250,000 / sq.cm	62500 pairs @ 30 Gbps → 1900 Tbps / sq.cm	Limited by thermal and coplanarity issues.
Stacked Silicon Carriers	Vertical connections @ 100 um pitch → 10,000 / sq.cm	2500 pairs @ 30 Gbps → 75 Tbps / sq.cm	Early demonstration only. Air cooled to < 117 W total.

Figure 6.36: Representative current and future high-end level 1 packaging.

3D Packaging Examples



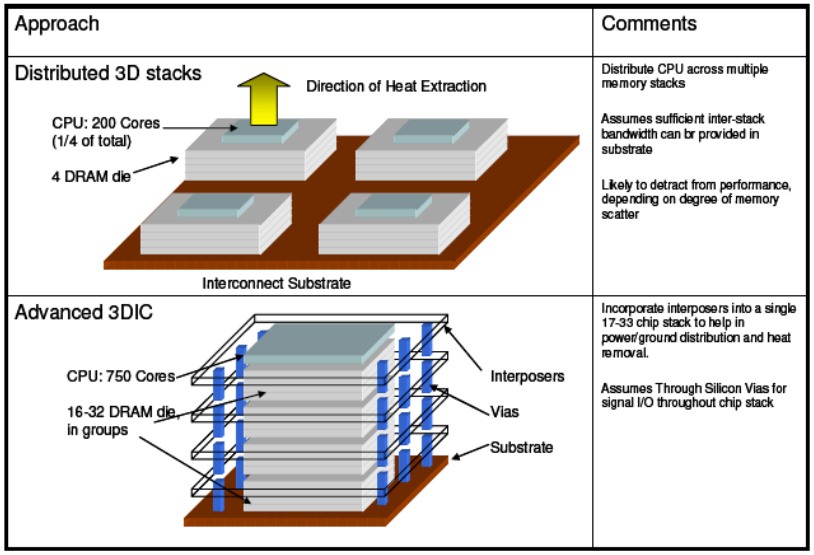


Figure 7.5: Potential directions for 3D packaging (A).

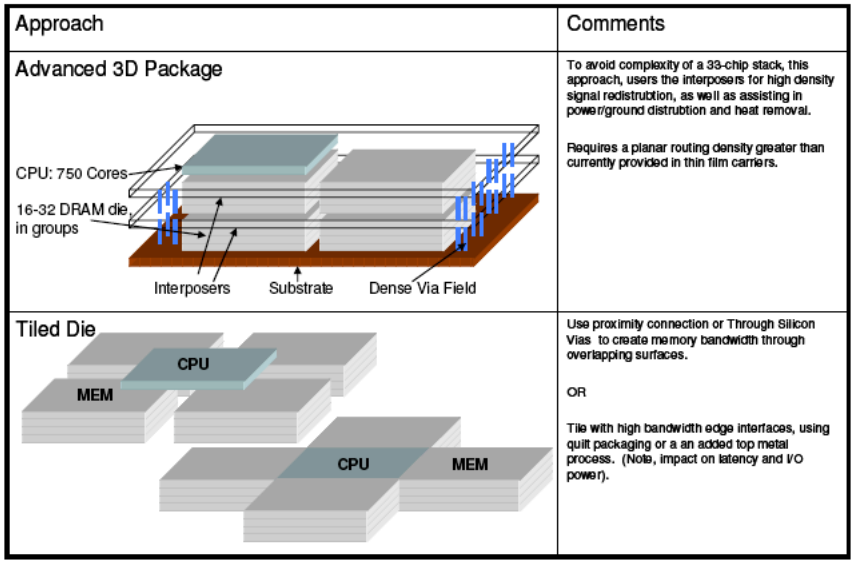
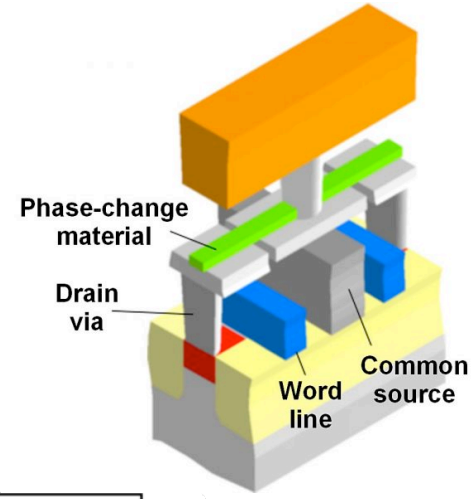


Figure 7.6: Potential directions for 3D packaging (B).

Secondary Storage Projections (Scratch at 25x , Archive at 200x)



	PB of Main Memory				
	0.006	0.5	3.6	50	300
Scratch Storage					
Capacity (EB)	1.2E-04	0.01	0.15	2	18
Drive Count	1.0E+01	8.3E+02	1.3E+04	1.7E+05	1.5E+06
Power (KW)	9.4E-02	7.8E+00	1.2E+02	1.6E+03	1.4E+04
Checkpoint Time (sec)	1.2E+03	1.2E+03	5.8E+02	6.0E+02	4.0E+02
Checkpoint BW (TB/s)	5.0E-03	4.2E-01	6.3E+00	8.3E+01	7.5E+02
Archival Storage					
Capacity (EB)	0.0012	0.1	7.2	100	600
Drive Count	1.0E+02	8.3E+03	6.0E+05	8.3E+06	5.0E+07
Power (KW)	9.4E-01	7.8E+01	5.6E+03	7.8E+04	4.7E+05

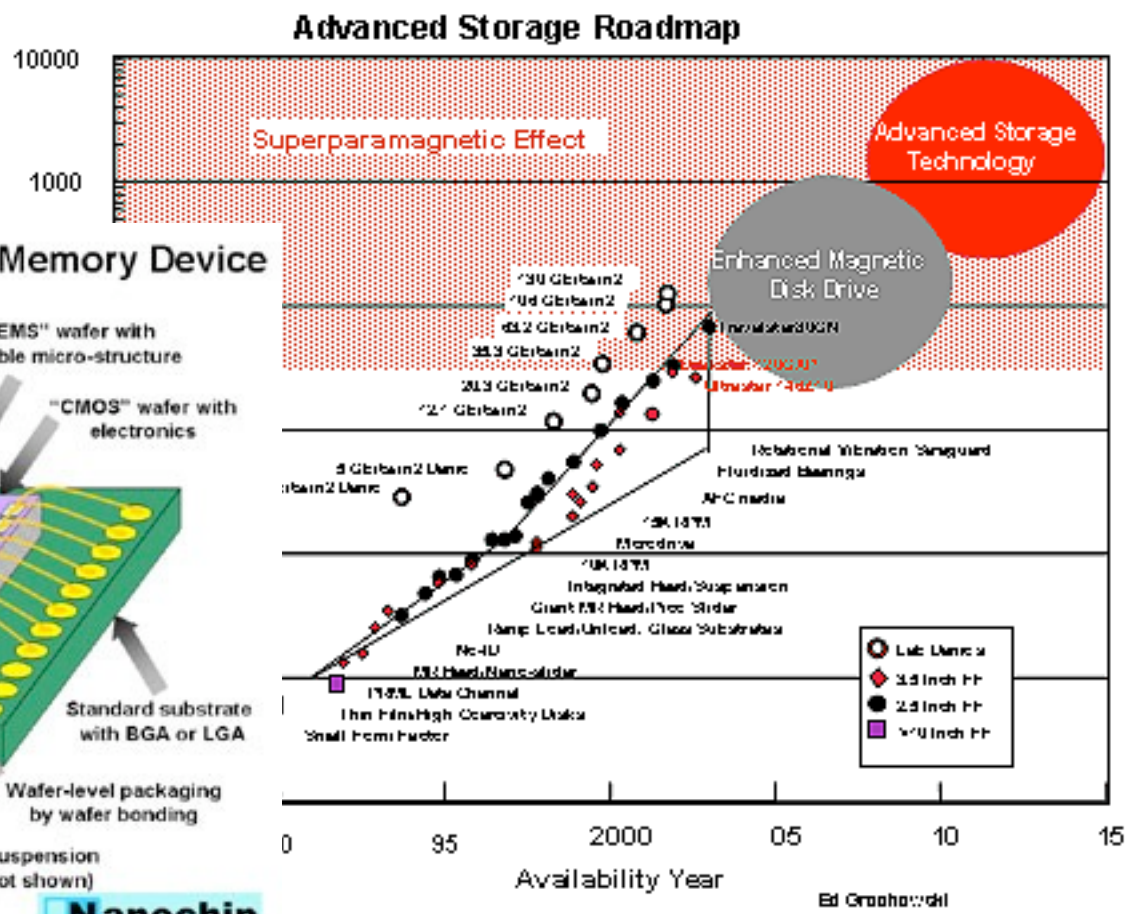
Table 7.1: Non-memory storage projections for Exascale systems.

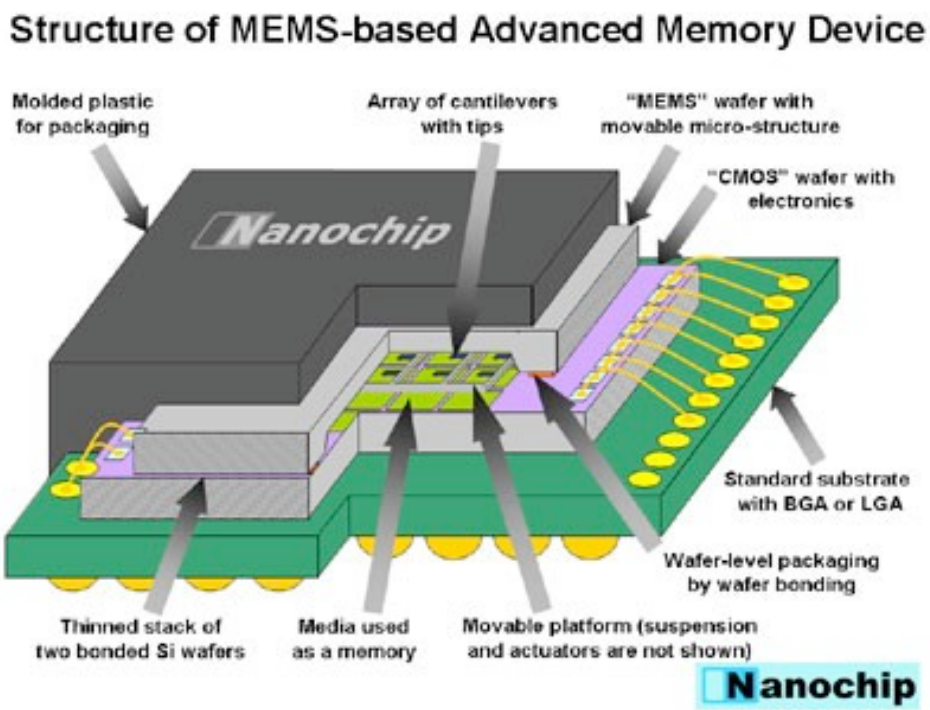
20,000 to > 1,000,000 disk drives for secondary storage

Phase change memory may replace disk in 2018 timeframe

Evolution of Storage

Year	Class	Capacity (GB)	RPM	B/W (Gb/s)	Idle Power(W)	Active Power (W)
2007	Consumer	1000	7200	1.03	9.30	9.40
2010	Consumer	3000	7200	1.80	9.30	9.40
2014	Consumer	12000	7200	4.00	9.30	9.40
2007	Enterprise	300	15000	1.20	13.70	18.80
2010	Enterprise	1200	15000	2.00	13.70	18.80
2014	Enterprise	5000				
2007	Handheld	60				
2010	Handheld	200				
2014	Handheld	800				





	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Chip Level Predictions															
Relative Max Power per Microprocessor	1.00	1.05	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
Cores per Microprocesor	2.00	2.52	4.00	5.04	6.36	8.01	10.09	12.71	16.02	20.18	25.43	32.04	40.37	50.85	64.07
Flops per cycle per Core	2.00	2.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	8.00	8.00	8.00	8.00	8.00	8.00
Flops per cycle per Microprocessor	4.00	5.05	16.00	20.17	25.44	32.04	40.37	50.85	64.07	161.43	203.40	256.29	322.92	406.81	512.57
Power Constrained Clock Rate	1.00	0.94	1.10	0.99	0.90	0.81	0.88	0.79	0.71	0.80	0.72	0.83	0.73	0.65	0.59
Relative Rpeak per Microprocessor	1.00	1.19	4.39	4.98	5.76	6.48	8.88	9.99	11.42	32.38	36.79	52.86	58.73	66.08	75.51
Actual Rpeak per Microprocessor	9.60	11.44	42.15	47.82	55.26	62.16	85.27	95.93	109.64	310.82	353.20	507.46	563.85	634.33	724.94
ITRS Commodity Memory Capacity Growth	1.00	1.00	1.00	2.00	2.00	2.00	4.00	4.00	4.00	8.00	8.00	8.00	16.00	16.00	16.00
Required Memory Chip Count Growth	1.00	1.19	4.39	2.49	2.88	3.24	2.22	2.50	2.86	4.05	4.60	6.61	3.67	4.13	4.72
Relative Grow												00	16.00	16.00	16.00
BW Scaled Relati												86	58.73	66.08	75.51
BW Scaled Relative												86	58.73	66.08	75.51
Simplistically												52	1.28	1.31	1.36
Fully Scaled												44	35.32	42.11	51.64
Simplistically S												38	45.98	50.26	55.42
Fully Scaled Relative Rpeak/Watt	1.00	1.06	1.24	1.36	1.53	1.31	1.32	1.77	1.71	1.36	1.40	1.19	1.66	1.57	1.46
Simplistically Scaled Rpeak/Watt	0.04	0.05	0.13	0.17	0.19	0.21	0.31	0.34	0.38	1.00	1.10	1.41	1.86	2.04	2.25
Fully Scaled Rpeak/Watt	0.04	0.04	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.06	0.06	0.05	0.07	0.06	0.06
Board and Rack Level Concurrency Predictions															
Maximum Sockets per Board	4	4	4	4	8	8	8	8	8	16	16	16	16	16	16
Maximum Boards per Rack	24	24	24	24	32	32	32	32	32	32	32	32	32	32	32
Maximum Sockets per Rack	96	96	96	96	256	256	256	256	256	512	512	512	512	512	512
Maximum Cores per Board	8	10	16	20	51	64	81	102	128	323	407	513	646	814	1025
Maximum Cores per Rack	192	242	384	484	1628	2050	2584	3254	4101	10331	13018	16402	20667	26036	32804
Maximum Flops per cycle per Board	16	20	64	81	204	256	323	407	513	2583	3254	4101	5167	6509	8201
Maximum Flops per cycle per Rack	384	484	1536	1936	6513	8201	10336	13018	16402	82650	104142	131218	165336	208285	262436
Board and Rack Level Power Predictions															
Max Relative Power per Rack	1	1	1	2	2	2	4	4	4	8	8	8	16	16	16
Simplistic Power-Limited Sockets/Rack	96	92	72	96	158	155	256	256	256	512	512	507	512	512	512
Fully Scaled Power-Limited Sockets/Rack	96	86	27	61	52	45	79	68	58	37	31	17	43	36	30
Simplistically Scaled Relative Rpeak per Rack	96	109	316	478	911	1001	2274	2558	2924	16577	18838	26788	30072	33831	38664
Fully Scaled Relative Rpeak per Rack	96	102	119	304	298	291	699	681	658	1197	1123	914	2554	2410	2246
System Predictions: Power Unconstrained, Gradual Increase in Affordable Racks to Max of 600															
Max Affordable Racks per System	155	200	250	300	350	400	450	500	550	600	600	600	600	600	600
Max Cores per System	23760	48444	9.8E+04	1.5E+05	5.7E+05	8.2E+05	1.2E+06	1.6E+06	2.3E+06	6.2E+06	7.8E+06	9.8E+06	1.2E+07	1.6E+07	2.0E+07
Max Flops per cycle per System	59520	96882	3.8E+05	5.8E+05	2.3E+06	3.3E+06	4.7E+06	6.5E+06	9.0E+06	5.0E+07	6.2E+07	7.9E+07	9.9E+07	1.2E+08	1.6E+08
Simplistically Scaled System Rpeak (GF)	1.0E+05	1.5E+05	5.4E+05	9.8E+05	2.2E+06	2.7E+06	7.0E+06	8.7E+06	1.1E+07	6.8E+07	7.7E+07	1.1E+08	1.2E+08	1.4E+08	1.6E+08
Fully Scaled System Rpeak (GF)	1.0E+05	1.4E+05	2.0E+05	6.2E+05	7.1E+05	7.9E+05	2.1E+06	2.3E+06	2.5E+06	4.9E+06	4.6E+06	3.7E+06	1.0E+07	9.9E+06	9.2E+06
System Power (MW)	2.5	3.2	4.0	9.7	11.3	12.9	29.0	32.3	35.5	77.4	77.4	77.4	154.8	154.8	154.8
System Predictions: Power Constrained to 20 MW															
Maximum Racks	155	200	250	300	350	400	310	310	310	155	155	155	78	78	78
Simplistically Scaled System Rpeak (GF)	1.E+05	1.E+05	5.E+05	1.E+06	2.E+06	3.E+06	5.E+06	5.E+06	6.E+06	2.E+07	2.E+07	3.E+07	2.E+07	2.E+07	2.E+07
Fully Scaled System Rpeak (GF)	1.E+05	1.E+05	2.E+05	6.E+05	7.E+05	8.E+05	1.E+06	1.E+06	1.E+06	1.E+06	1.E+06	1.E+06	1.E+06	1.E+06	1.E+06

Cores per Processor grows from 6 in 2009 to > 64 in 2020
Operations per clock per clock per processor from 16 to 512
Rack count grows from 200 to over 600

Figure 7.10: Heavy node strawman projections.

Aggressive Strawman

Level	What	Perf	Power	RAM
FPU	FPU, regs,. Instruction-memory	1.5 Gflops	30mW	
Core	4FPUs, L1	6 Gflops	141mW	
Processor Chip	742 Cores, L2/L3, Interconnect	4.5 Tflops	214W	
Node	Processor Chip, DRAM	4.5Tflops	230W	16GB
Group	12 Processor Chips, routers	54Tflops	3.5KW	192GB
rack	32 Groups	1.7 Pflops	116KW	6.1 TB
System	583 racks	1 Eflops	67.7MW	3.6PB

Table 7.3: Summary characteristics of aggressively designed strawman architecture.

~200,000 nodes each with ~750 cores = ~150M cores
Memory is 16GB for a ~5 Tflop node with ~1000 threads
Total system memory is ~ 4PB.

Systems Scaling Projections

Begin Full System Delivery (Yr)	2004	2007	2012	2015	2019
Design Parameters	**BG/L**	**BG/P**	**25PF**	**300PF**	**1200PF**
Cores / Node	2	4	8-24	32-64-128	96-128-500
Clock Speed (GHz)	0.7	0.85	1.6-4.1	2.3-4.8	2.8-6.0
Flops / Clock / Core	4	4	8-32	8-32	16-64
Nodes / Rack	1024	1024	100-1024	256-1024	256-1024
Racks / Full System Config	64	72	128-350	128-400	256-400
MB RAM/core	256	512	1024-4096	1024-4096	1024-4096
Total Power	2.5MW	4.8MW	8MW-20MW	20MW-50MW	30MW-80MW
Flops / Node (GF)	5.6	14	128-640	640-2000	2000-6000
Flops / Rack (TF)	5.7	14	200-400	400-1200	1600-4800
LB Concurrency	5.E+05	1.E+06	1M-2M	10M-100M	400M-1000M
Full System					
Total Cores (Millions)	0.13	0.3	.3M-1.2M	1M-10M	4M-200M
Total RAM (TB)	33.6	151	2,000-4,400	3,000-10,000	5,000-25,000
Total Racks	64	72	128-350	128-400	256-400
Peak Flops System (PF)	**0.37**	**1**	**25**	**300**	**1200**

The Bottom Line

- Levels of concurrency ($10^6 \Rightarrow 10^9$) 1000x
- Clock rate of Core (1-4 GHz $\Rightarrow$ 1-4 GHz) 1x
- RAM per Core (1-2GB now to <1GB) <1x
- Total Number of cores (200K $\Rightarrow$ 100M) 500x
- Number of cores per node (8 $\Rightarrow$ >64-512) >8x-64x
- Aggressive Fault Management in HW and SW
- I/O channels ($>10^3 \Rightarrow 10^5$) >100x
- Power Consumption (10MW $\Rightarrow$ 40MW-150MW) 4x-16x
- Programming Model (MPI $\Rightarrow$ MPI + X)

Parallel Programming Models: Twenty Years and Counting

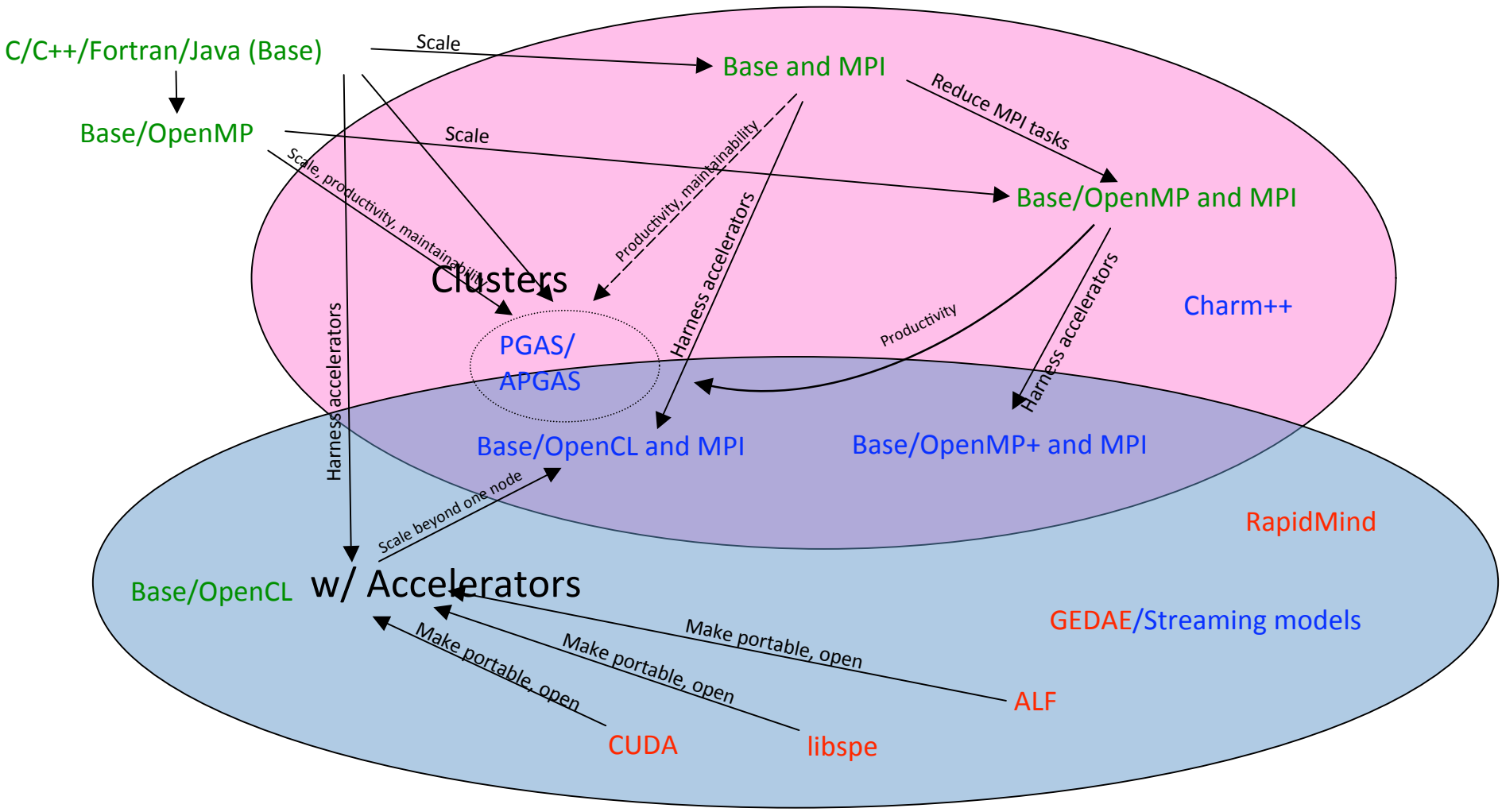
- In large-scale scientific computing today essentially all codes are message passing based. Additionally many will use some form of multithreading on SMP or multicore nodes.
- Multicore is challenging programming models but there has not yet emerged a dominate model to augment message passing
- There is a need to identify new hierarchical programming models that will be stable over long term and can support the concurrency doubling pressure

Quasi Mainstream Programming Models

- C, Fortran, C++ and MPI
- OpenMP, pthreads
- (CUDA, RapidMind, Cn) → OpenCL
- PGAS (UPC, CAF, Titanium)
- HPCS Languages (Chapel, Fortress, X10)
- HPC Research Languages and Runtime
- HLL (Parallel Matlab, Grid Mathematica, etc.)

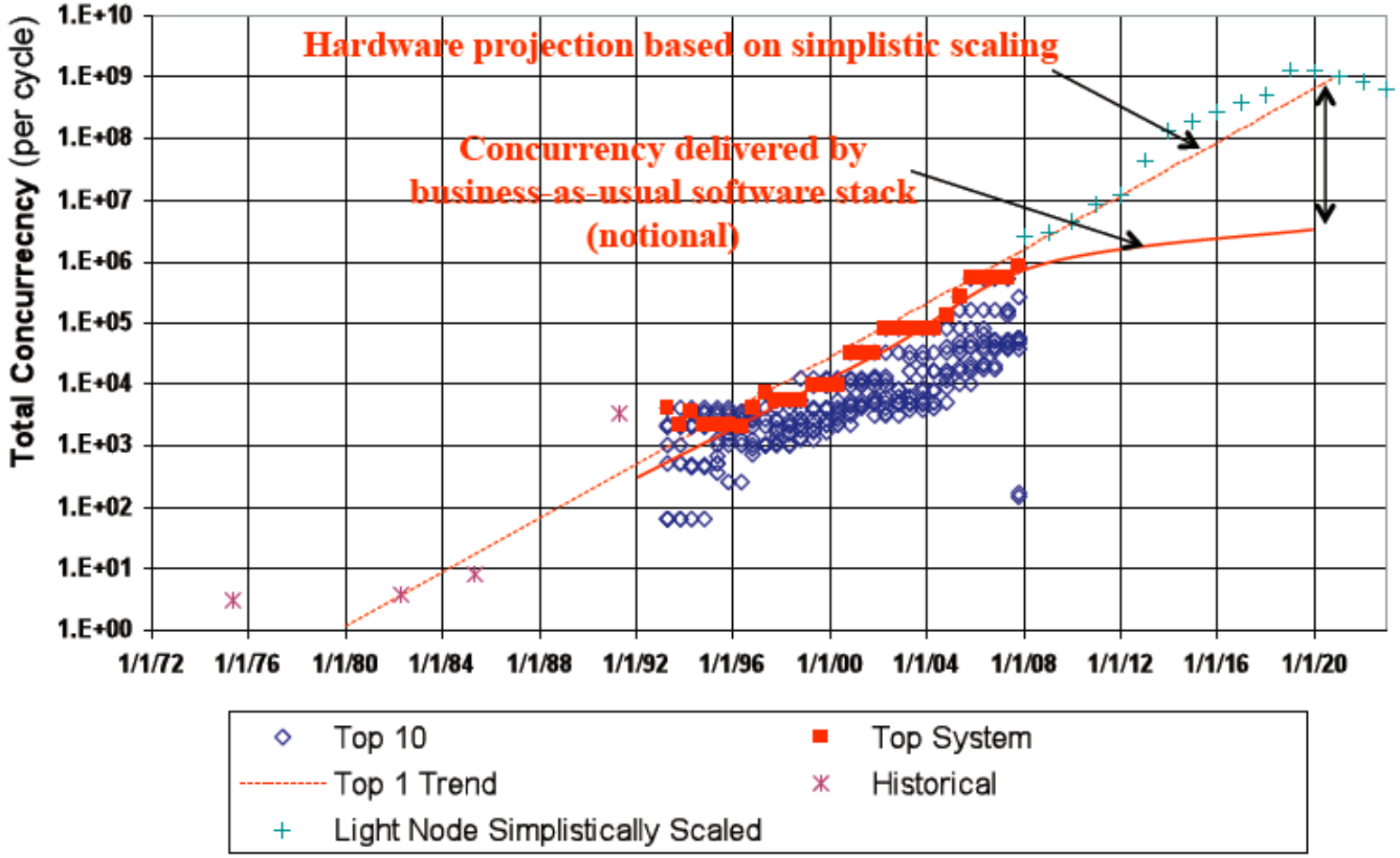
Potential Migration Paths Programming Models

Green: open, widely available*
Blue: somewhere in between
Red: proprietary
*OpenCL availability predicted



The Exascale Software Challenge

How to bridge the gap between Simplistically Scaled projections and reality of application software?



"ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems", P.Kogge et al

From Vivek Sarkar SC08 Exascale Workshop

Concurrency Challenge

- **How will 1000x increase in concurrency be delivered to Extreme Scale Systems?**
  - Weak Scaling by increase in data size, W_S
  - Weak Scaling by increase in computation per datum, W_T
  - Strong Scaling by increasing parallelism in fixed computation, S

- **Over Provisioning required for Latency Hiding, L**
  - L is determined by latency hiding requirements --- memory access time, fraction of memory operations (parameterized by memory hierarchy level
  - Assume L ~ 100x

- **How much additional parallelism must come from software, relative to today's applications?**
  - ***W_S*W_T*S needs to be > 10^5***

From Vivek Sarkar SC08 Exascale Workshop

Concurrency Challenge (contd)

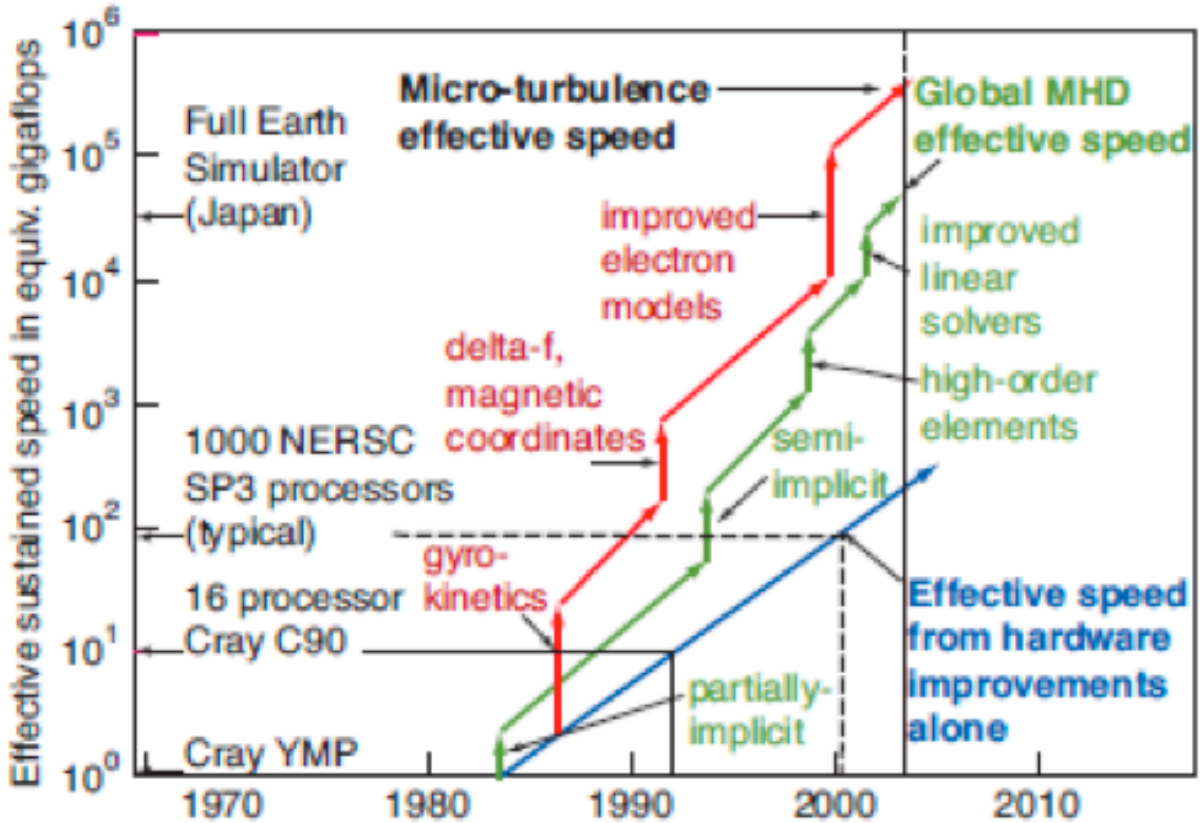
- **W_S is limited by DRAM cost and memory-computation ratio**
  - Petascale ~ 1 byte/FLOPS
  - Exascale ~ 0.01 bytes/FLOPS ??
  - Improving memory-computation balance is expensive
    - 50-50 cost balance dictates 0.004 bytes/FLOPS
    - 90-10 cost balance dictates 0.04 bytes/FLOPS
  - Assume W_S ~ 10x

- **W_T is limited by application domain and algorithm**
  - Assume W_T ~ 10x

➔ **Software must deliver S ~ 1000x increase in Strong Scaling to use full capability of Extreme Scale systems**

From Vivek Sarkar SC08 Exascale Workshop

Algorithms and models also yield solutions



Magnetic fusion energy "effective speed" increases
Source: A Science-Based Case for Large-Scale Simulation (SCaLeS) Volume 2

A balanced R&D portfolio includes all three:

-- faster hardware

-- better mathematical models

-- improved computational algorithms

U.S. DEPARTMENT OF ENERGY

	Exascale System Class				
Characteristic	Exaflops Data Center	20 MW Data Center	Department	Embedded A	Embedded B
Top-Level Attributes					
Peak Flops (PF)	9.97E+02	303	1.71E+00	4.45E-03	1.08E-03
Cache Storage (GB)	3.72E+04	11,297	6.38E+01	1.66E-01	4.03E-02
DRAM Storage (PB)	3.58E+00	1	6.14E-03	1.60E-05	1.60E-05
Disk Storage (PB)	3.58E+03	1,087	6.14E+00	0.00E+00	0.00E+00
Total Power (KW)	6.77E+04	20,079	116.06	0.290	0.153
Normalized Attributes					
GFlops/watt	14.73	14.73	14.73	15.37	7.07
Bytes/Flop	3.59E-03	3.59E-03	3.59E-03	3.59E-03	1.48E-02
Disk Bytes/DRAM Bytes	1.00E+03	1.00E+03	1.00E+03	0	0
Total Concurrency (Ops/ Cycle)	6.64E+08	2.02E+08	1.14E+06	2968	720
Component Count					
Cores	1.66E+08	50,432,256	2.85E+05	742	180
Microprocessor Chips	223,872	67,968	384	1	1
Router Chips	223,872	67,968	384	0	0
DRAM Chips	3,581,952	1,087,488	6,144	16	16
Total Chips	4,029,696	1,223,424	6,912	17	17
Total Disk Drives	298,496	90,624	512	0	0
Total Nodes	223,872	67,968	384	1	1
Total Groups	18,656	5,664	32	0	0
Total racks	583	177	1	0	0
Connections					
Chip Signal Contacts	8.45E+08	2.57E+08	1.45E+06	2,752	2,752
Board connections	1.86E+08	5.65E+07	3.19E+05	0	0
Inter-rack Channels	2.35E+06	7.14E+05	8,064	0	0

Table 7.10: Exascale class system characteristics derived from aggressive design.