Findings and Recommendations from the Petaflops Workshop Series

Thomas Sterling

NASA Jet Propulsion Laboratory

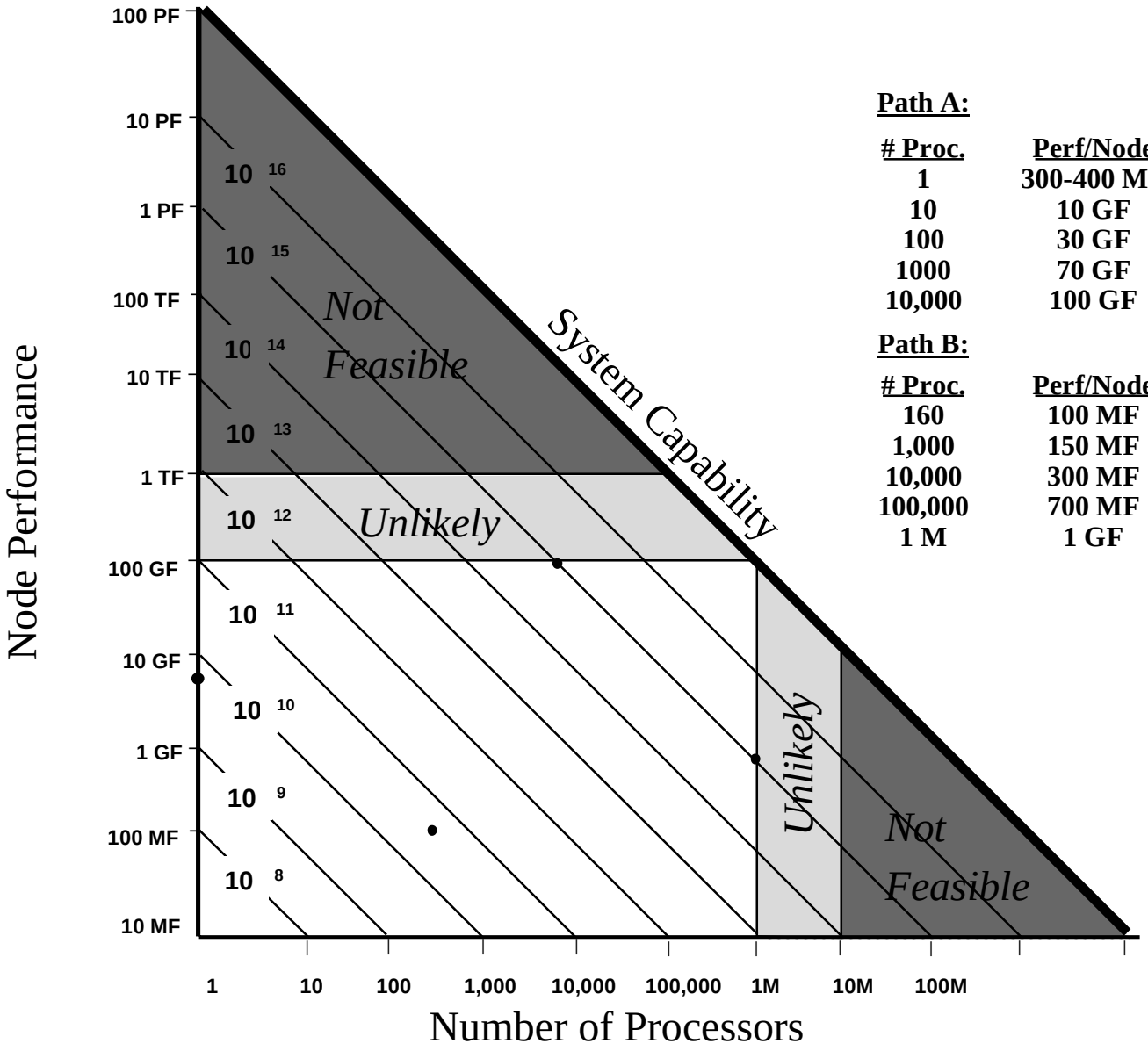
&

California Institute of Technology

June 24, 1997

JPL

Path(s) to PetaFLOPS



Path A:

# Proc.	Perf/Node
1	300-400 MF
10	10 GF
100	30 GF
1000	70 GF
10,000	100 GF

Path B:

# Proc.	Perf/Node
160	100 MF
1,000	150 MF
10,000	300 MF
100,000	700 MF
1 M	1 GF

Workshops on Petaflops Computing

- Goal
  - Determine feasibility, time frame, and means of achieving petaflops performance
- Focus Issues
  - Technology
  - Architecture
  - System Software
  - Applications
  - Algorithms

Petaflops Workshops

- First Workshop on Enabling Technologies for Petaflops Computing (Pflops-1): February, 1994
- First Petaflops Frontier Workshop:(TPF-1): February, 1995
- Petaflops Applications Summer Study (Bodega Bay): August, 1995
- Petaflops Architecture Workshop (PAWS): April, 1996
- Petaflops System Software Workshop (PetaSoft): June, 1996
- Second Petaflops Frontier Workshop (TPF-2): October, 1996
- System Software Mini-Workshop (MiniSoft): February, 1997
- Petaflops Algorithms Workshop (PAL): April, 1997
- 2nd Workshop on Enabling Technologies for Petaflops Computing (Pflops-2): January, 1998

Sponsoring Agencies

- BMDO
- DARPA
- DOE
- NASA
- NSA
- NSF

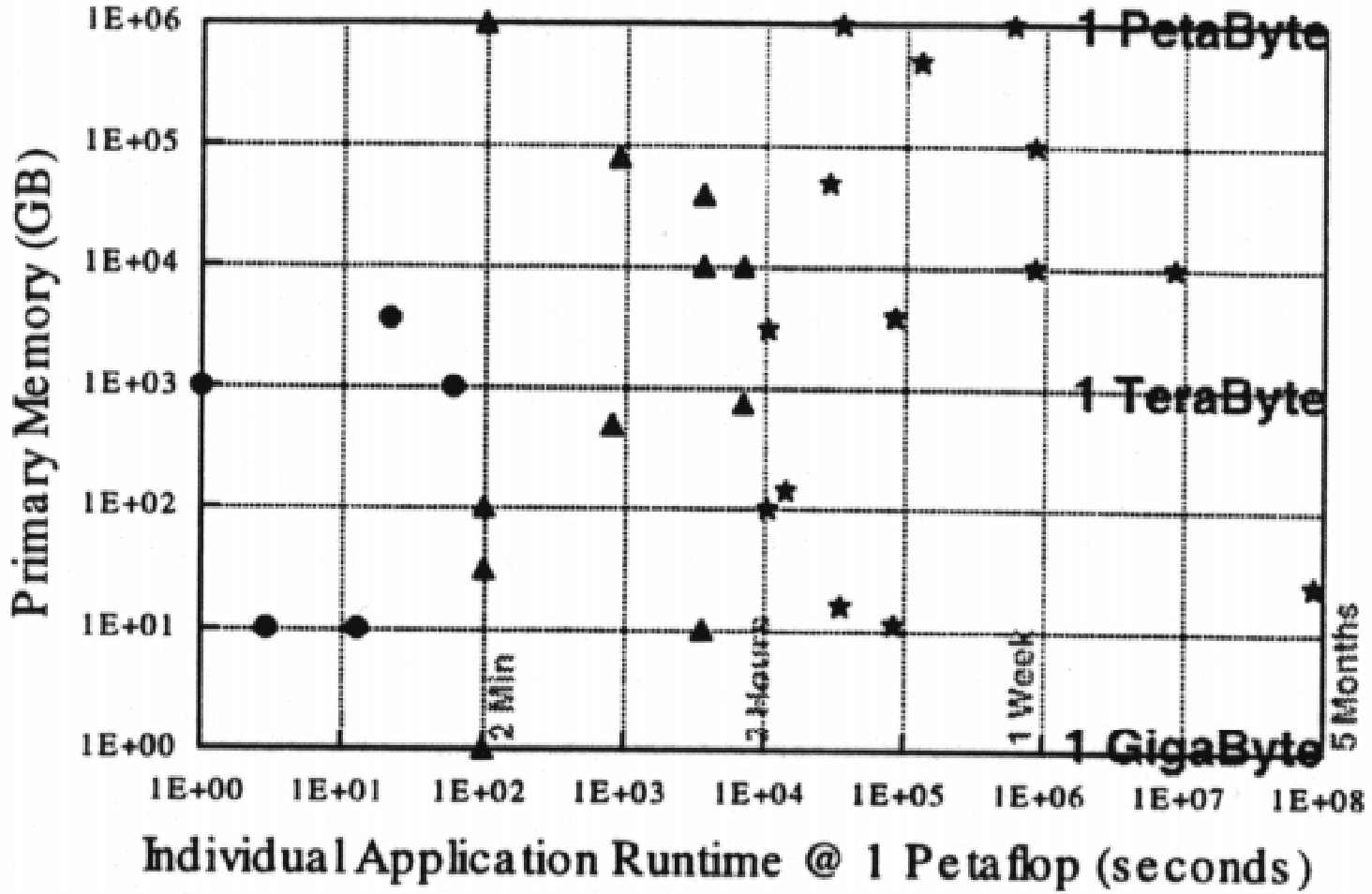
Strategic Findings

- Petaflops capability is essential for many important tasks in science, industry, defense, and commerce
- Petaflops systems will be feasible by 2015 to 2020
- Petaflops performance may be achievable by 2007 or earlier through alternative approaches
- For accelerated path, paradigm shift required
- Cost, power, and efficiency factors dominate
- Innovative methods and structures critical to generality, usability, efficiency, and cost effectiveness

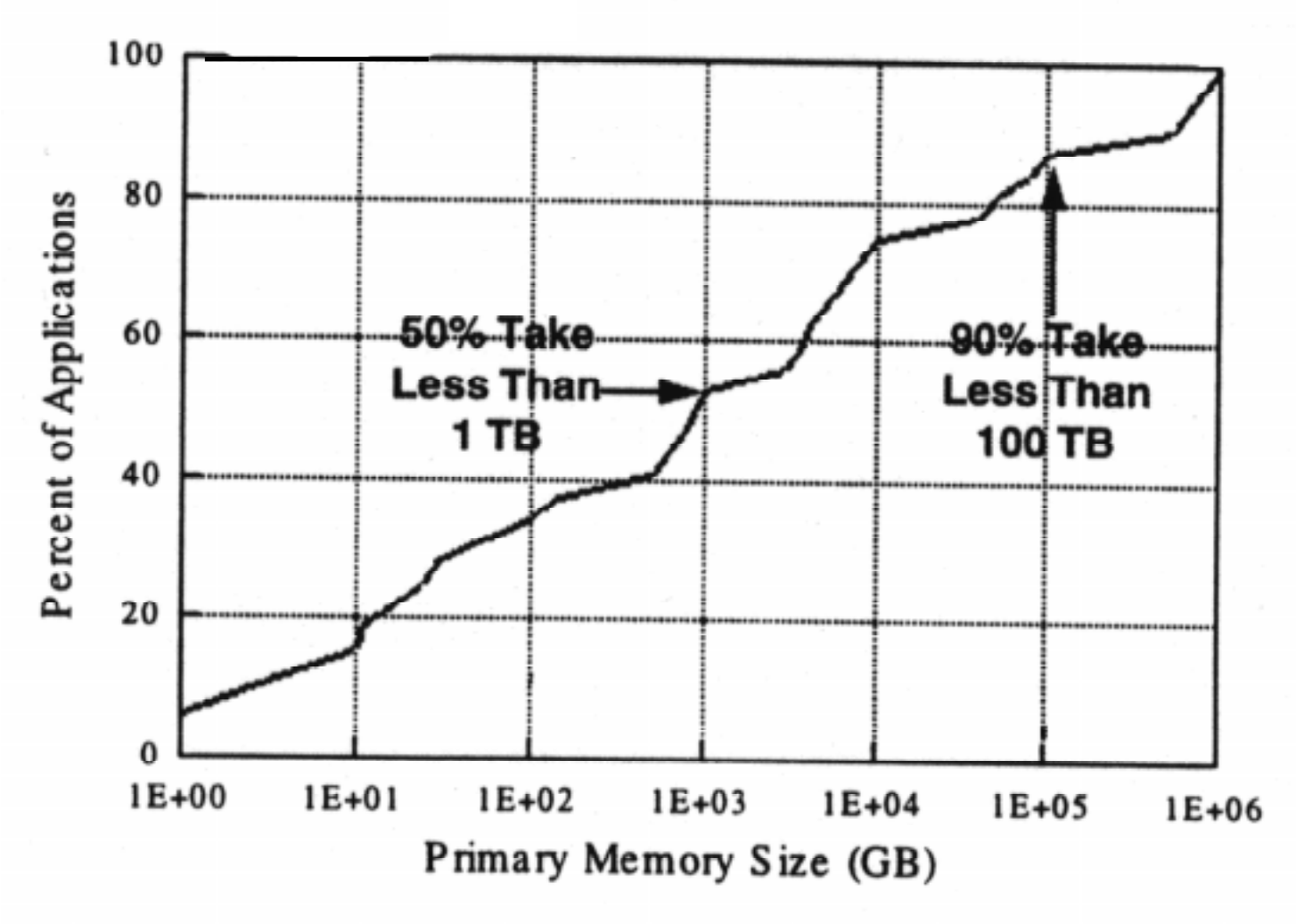
Applications for Petaflops Computers

- Nuclear weapons stewardship
- Cryptology and digital signal proc.
- Satellite data processing
- Climate and environmental modeling
- 3-D protein molecular reconstructions
- Severe storm forecasting
- Design of advanced aircraft
- DNA sequence matching
- Molecular nanotechnology
- Large-scale economic modeling

How About Real Applications?



What Are Realistic Storage Sizes?



The NSF Point Design Studies

- *A Flexible Architecture for Executing Component Software at 100 Teraops* (University of Illinois at Urbana-Champaign)
- *Pursuing a Petaflop: Point Designs for 100 TF Computers Using PIM Technologies* (Notre Dame)
- *Architecture, Algorithms and Applications for Future Generation Supercomputers* (University of Minnesota)
- *Design Studies on Petaflops Special-purpose Hardware for Astrophysical Particle Simulations* (Drexel University, University of Tokyo, University of Illinois at Urbana-Champaign, and Princeton University)
- *Hybrid Technology Multithreaded Architecture* (California Institute of Technology, University of Delaware, and State University of New York at Stony Brook)
- *Hierarchical Processors and Memory Architecture for High Performance Computing* (Purdue University and Northwestern University)
- *The Illinois Aggressive Cache-only Memory Architecture Multiprocessor, I-ACOMA* (University of Illinois at Urbana-Champaign)
- *A Scalable-Feasible Parallel Computer Implementing Electronic & Optical Interconnections for 156 Teraops Minimum Performance* (New Jersey Institute of Technology and Wayne State University)

Challenges

- Peak Performance
  - Aggregate Compute Resources
  - Cost
  - Power consumption
- Sustained Performance
  - Latency
  - Overhead
  - Starvation
  - Contention
  - Generality (*Qness)*
  - Programmability

Architecture Concepts

- Superconducting processor logic
- Direct mapping of physical dataflow paths to algorithm kernels
- Exposing intrinsic memory structures to logic
- Compiler/library visible datapath resident reconfigurable logic
- Parallel speculative execution support for runtime address disambiguation
- Multithreading
- Aggressive prefetching into memory hierarchy
- COMA
- Ensemble of non-uniform granularity processors to beat Amdahl’s Law
- Hierarchical topologies with varying bandwidth to match
- Algorithms communications
- Free space full interconnection with optics

Basic SIA Roadmap CMOS Trends

Characteristic (SIA Pg. #)	**Units**	**1995**	**1998**	**2001**	**2004**	**2007**	**2010**
Feature Size(11)	μm	0.35	0.25	0.18	0.13	0.1	0.07
Vdd(14)	volts	3.3	2.5	1.8	1.5	1.2	0.9
DRAM							
Chip Capacity	MB	8	32	128	512	2,048	8,192
Chip Size(12)	mm^2	190	280	420	640	960	1400
Density	MB/cm^2	4	11	30	80	213	585
Chip Cost	Rel. to 1995	1	1.65	2.82	3.76	7.53	12.05
\$/MB	Rel. to 1995	1	0.41	0.18	0.06	0.03	0.01
High Performance Microprocessor Logic Based Chips							
Transistors/Chip(16)	MT	12	28	64	150	350	800
Chip Size(B2)	mm^2	250	300	360	430	520	620
Density	MT/cm^2	5	9	18	35	67	129
Clock: μP(12)	MHz	300	450	600	800	1000	1100
Clock: DSP(46)	MHz	400	600	800	1100	1500	1900
SRAM Cache Density(11)	MB/cm^2	2	6	20	50	100	300
Cost/Transistor (B2)	millicents	1	0.5	0.2	0.1	0.05	0.02
Chip Cost	Rel. to 1995	1	1.17	1.07	1.25	1.46	1.33
ASIC Logic Chips							
Transistors/Chip	MT	9	26	53	108	275	560
Chip Size(B2)	mm^2	450	660	750	900	1100	1400
Logic Density(B2)	MT/cm^2	2	4	7	12	25	40
Clock(B2)	MHz	150	200	300	400	500	625
Minimum Chip Cost	Rel. to 1995 μP	0.75	1.1	0.88	0.9	1.15	0.93
NRE Chip Cost	\$/volume	27	26	26	32	55	56

Natural Evolutionary Path

- Approaches:
  - Pure COTS
    - e.g., Beowulf-class PC clusters,
      - NOW/COW
  - COTS plus
    - e.g., CRI T3E,
      - HP Convex SPP-2000

	1998	**2010**
Cost	$20,000M	$500M - $1000M
Power	1300 MVA	30 MVA

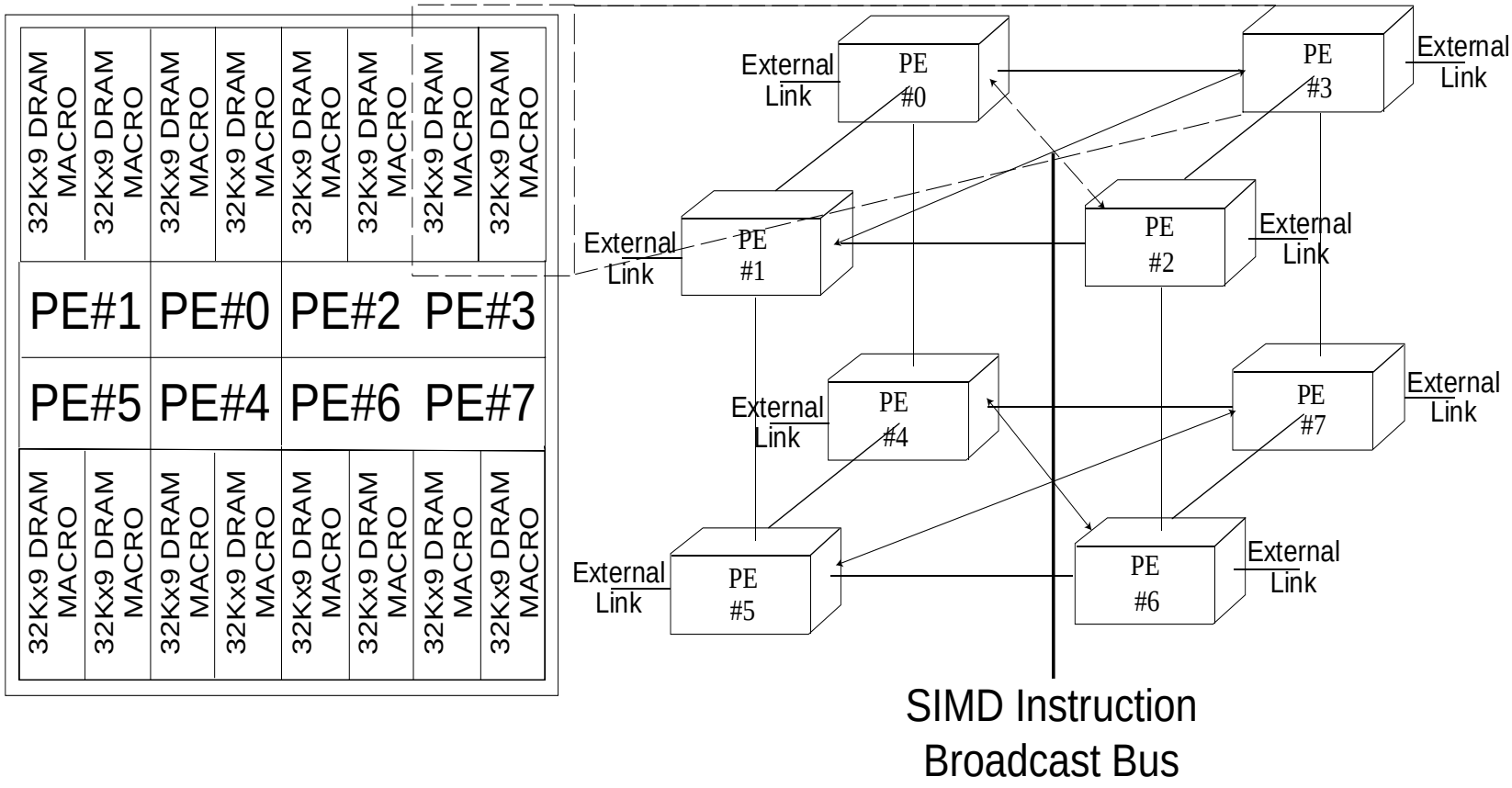
Particle # of Largest Simulation Feasible & Grape-6 Time Advantage†

Problem Area	1995 General Purpose Supercomputer	1995 GRAPE-4	2000 General Purpose Supercomputer	2000 GRAPE-6	GRAPE-6 Time Advantage
Planet Formation & Rings	5×10^3	5×10^4	5×10^4	10^6	7 years
Globular Cluster Evolution	10^4	5×10^4	5×10^4	5×10^5	10 years
Black Hole Binary in Galactic Nucleus	10^5	10^6	10^6	3×10^7	10 years
Galaxy Evolution & Interactions	10^6	2×10^6	3×10^7	10^8	3 years
Galaxy Evolution & Interactions with SPH	2×10^5	10^6	5×10^6	10^7	2 years
Large Scale Structure & Galaxy Formation	3×10^7	3×10^7	5×10^8	5×10^8	0 years**
Large Scale Structure & Galaxy Formation with SPH	4×10^6	3×10^6	10^8	3×10^8	3 years

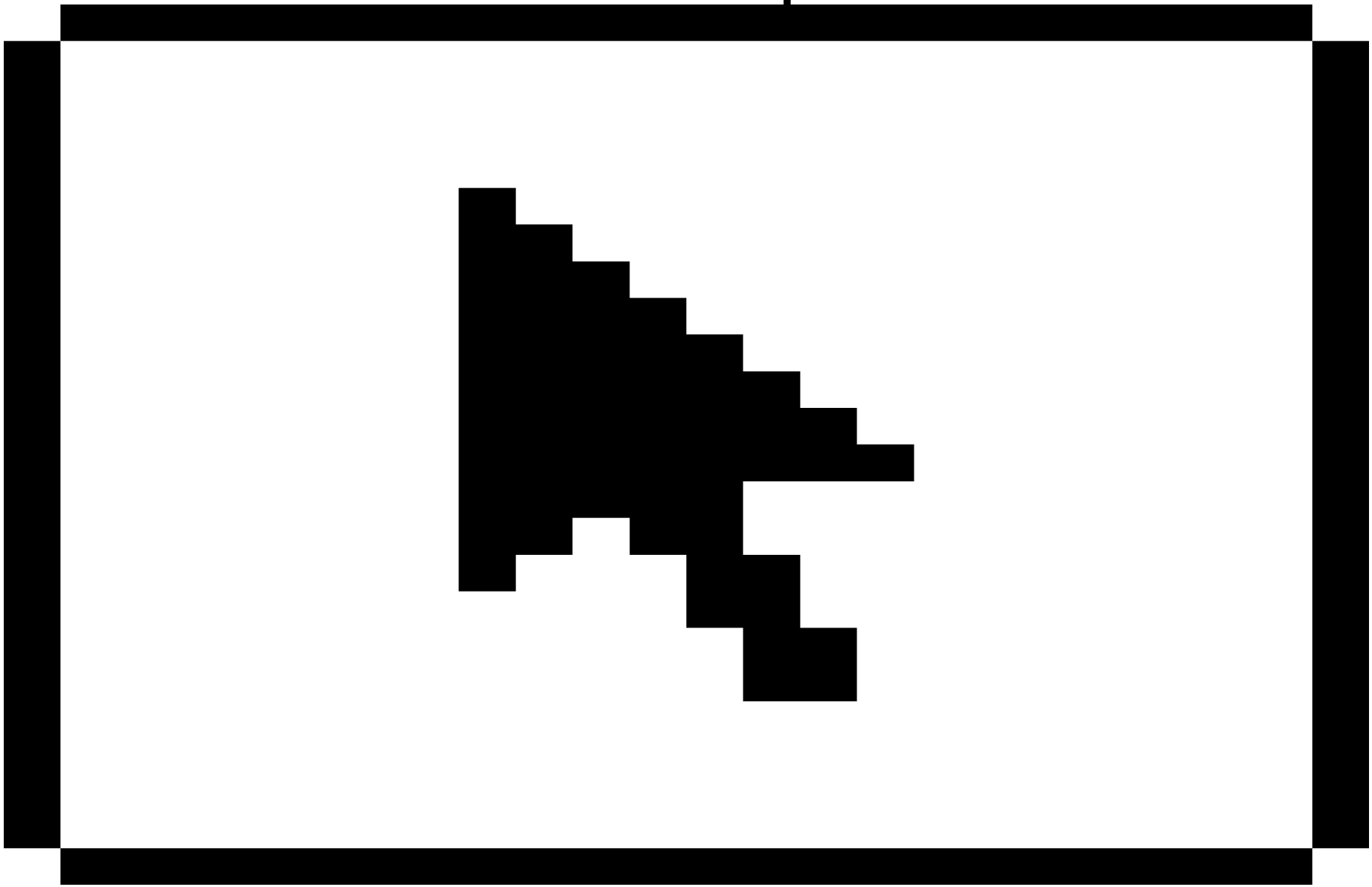
† All numbers are rough estimates for initial evaluation purposes only

** Assumes RAM limited calculation

Point Designs Using PIM Technologies



SIA Projections for PIM Technology to a 100 Teraflops Machine



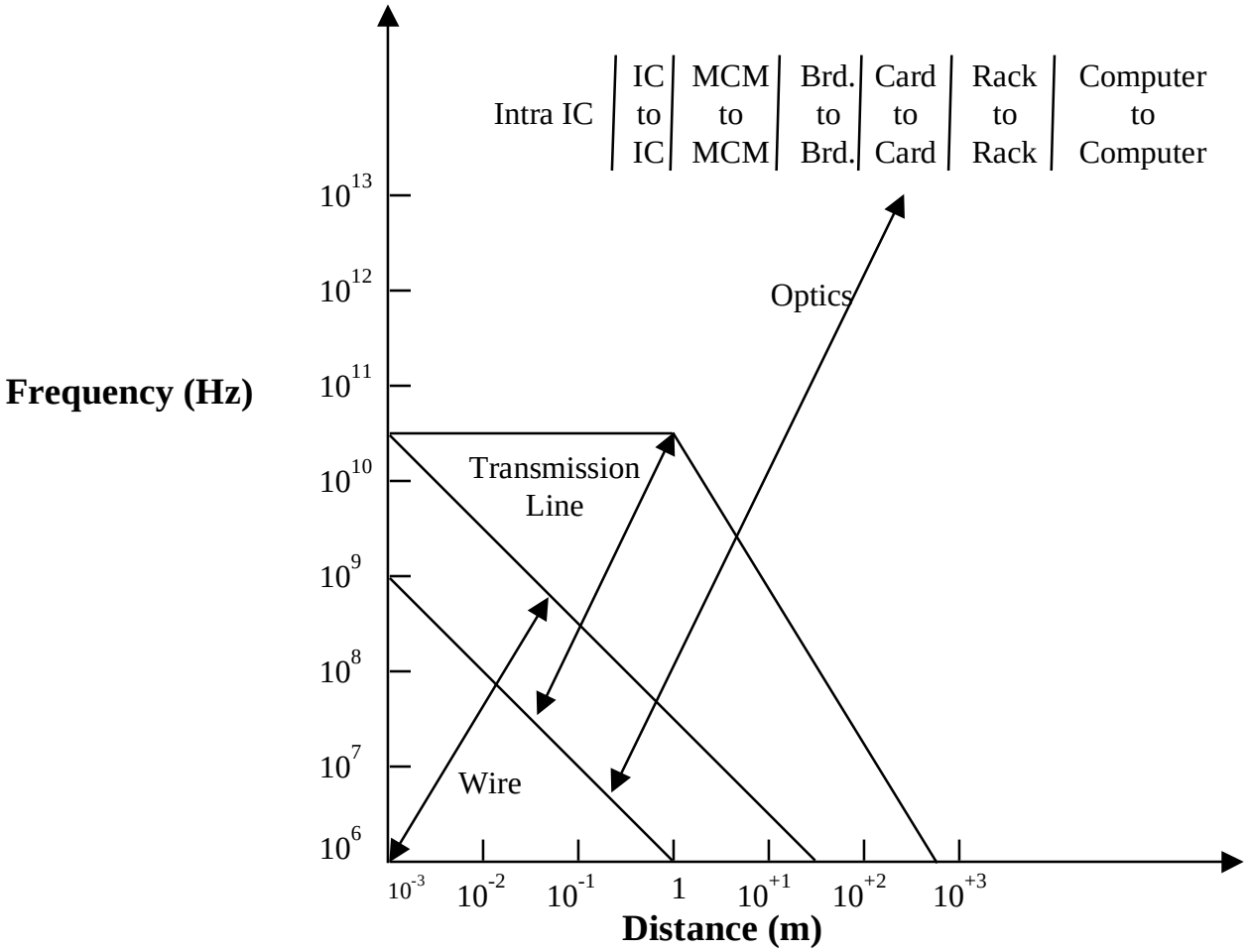
RSFQ TECHNOLOGY ROADMAP

Technology Parameters	HYPRES upgrade	SUNY upgrade	VLSI (shunted)	VLSI (unshunted)
Year	**1998**	**2001**	**2004**	**2007**
Josephson junction size (μm)	3.50	1.50	0.80	0.50
Logic circuit density (Kgates/cm^2)	10	30	100	1,000
Josephson current density (k/Vcm^2)	1	6.5	20	50
Specific capacitance (aF/μm^2)	45	60	67	75
I_cR_n product (mV)	0.3	0.6	1	1.5
SFQ pulse duration τ (ps)	4	2	1.2	0.8
Clock frequency f_{max} (GHz)	150	300	500	700
Speed of LSI circuits (GHz)	30	60	100	150
Average power (μw/gate)	0.03	0.06	0.1	0.15
Cost per junction (millicents)	100	30	10	1

Optical Holographic Storage

- Potential of terabyte-scale storage
- Very low power
- Very high bandwidth
  - 100 Gbps
- Photorefractive
- Spectral hole burning
- Near term technologies

Optics: The Preferred Interconnect



Optics: The preferred interconnect technology for the higher frequency and longer distance applications [Feldman:88a], [Tsang:90a].

Emerging Capabilities for Guided Optical Interconnects

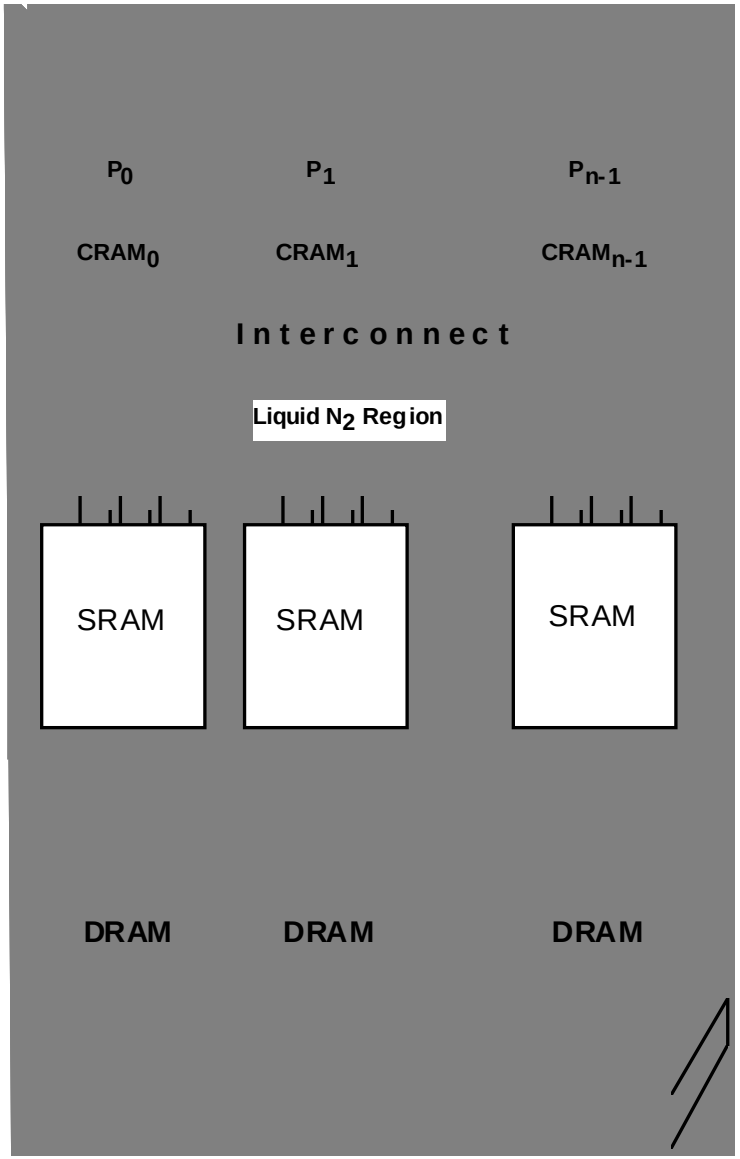
	Example	Speed	Power	Technology
Current Capability	HP and AT&T SONET	100 Mb/s 2.5 Gb/s	1-2 Watts 10 Watts	LED* LD*
10-Year Projection		100 Gb/s	1 Watt	LD (WDM or TDM)*
20-Year Projection		1 Tb/s	10 Watts	LD (WDM and TDM)

* LED (Light Emitting Diode), LD (Laser Diode), WDM (Wavelength Division Multiplexing), TDM (Time Division Multiplexing)

HTMT Concepts and Approach

- Deep memory hierarchy to match very high speed logic with high capacity storage
- Employ multi-stage multithreaded mechanisms for latency management
- Smart storage performs memory oriented operations
  - establish new processor/memory relationship
  - define Memory Instruction Set Architecture (MISA)

HTMT Architecture



Multistage Multithreaded Execution Model

- Extend latency hiding of multithreading
- Hierarchy of logical threads
  - Delineates threads and thread ensembles
  - Action sequences, state, and precedence constraints
- Fine grain single cycle thread switching
  - Processor level, hides pipeline and time of flight latency
- Coarse grain context “percolation”
  - Memory level, in memory synchronization
  - *Ready* contexts move toward processors, *pending* contexts towards big memory

System Software

- Importance of software
- Need for explicit, low-level control of hardware
- Need for a layered software architecture
- Existence of fundamental systems software scaling problems
- Opportunities for immediate software effort
- Interrelationship of hardware and low-level software

Algorithms

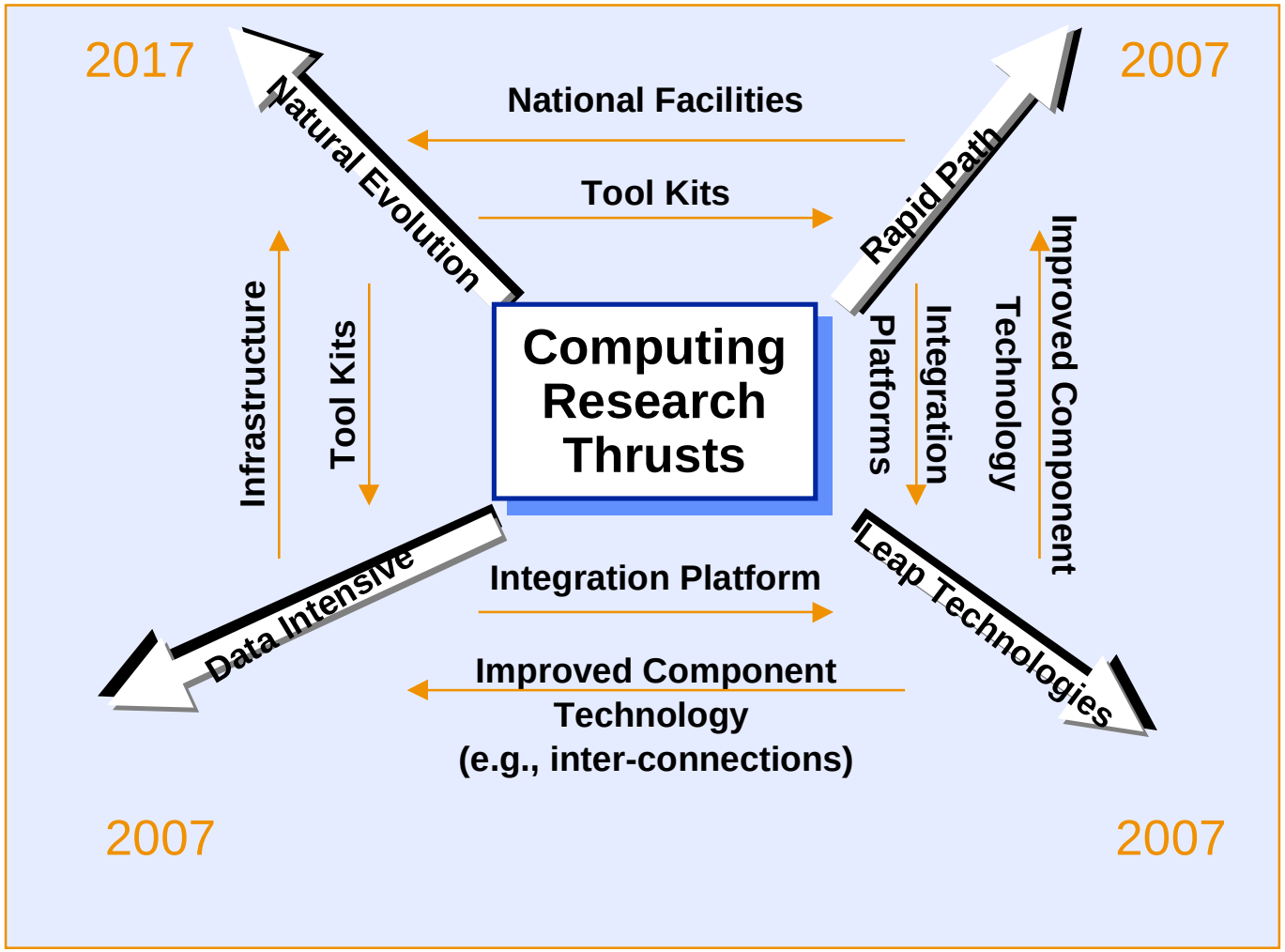
- Concurrency
- Data locality
- Latency and bandwidth
- Numerical accuracy
- New algorithmic approaches
- New languages and constructs
- Numerical library routines

Follow-on Investigations

- Pflops-1
  - Petaflops Applications Summer Study
  - PAWS
  - PetaSoft
- Petaflops Applications Summer Study
  - NSF Point Design Study
  - PAL Workshop
- PetaSoft
  - MiniSoft
- PAWS (Architecture Types and Exotic Technologies)
  - COTS:
  - HT + PIM + MT:
    - HTMT detailed study (sponsors - NSA, NASA, and DARPA)
    - Multi-stage Multithreading
  - SPD: Grape-6

    SPH SPD workshop (NASA sponsored)

Directions Toward Petaflops Computing



Major Technical Findings

- Petaflops applications vary widely in memory use
  - Petabyte memory for data intensive applications
  - 3/4 root (32 Terabytes) for *3-D+T* simulations
  - Small memory (<1 Terabytes) for narrow class of problems
- Latency management critical, requiring aggressive techniques in avoidance and hiding for efficiency
- Parallelism of 1,000,000 or more concurrent threads
- Bandwidth is primary obstacle to effective operation
- Architecture, software, and algorithms will differ from today's methods in response to these factors

Major Technical Findings (cont.)

- Exotic technologies offer aggressive advantage
  - Advanced Semiconductor
  - Superconducting RSFQ devices
  - Holographic memory
  - Optical interconnect
- Alternative architecture approaches
  - COTS derived clustered processor nodes
  - Processor in Memory (PIM)
  - Hybrid technology with deep memory hierarchy
  - Special purpose devices employing systolic structures
- Architecture advances required in processor design, aggressive latency control, and interconnection

• Major Technical Findings (cont.)

- System software resource management scalability identified as critical to utility
- Conflicting requirements challenge software design
  - expose hardware characteristics while supporting abstractions
  - emphasize reuse while exploiting specialized mechanisms
  - novel languages while supporting legacy code
- Hardware and low level software design interrelated
  - hardware provides efficient mechanisms
  - software establishes operation policies for application needs

Open Issues

- Can orders-of-magnitude latency be managed?
- What will the computer languages of the petaflops era look like?
- Should processor granularity be fat and few, or lean and unlimited?
- Can exotic technologies really outpace the CMOS stampede?
- Is it possible for system software of the future to turn a large pile of nodes back into a single computer?
- Is parallelism ubiquitous in the universe?

Recommendations

- Conduct detailed design and simulation studies of promising petaflops architectures
  - COTS technology based parallel computing
    - e.g., Pure COTS: Pile of PCs;
      - Tightly coupled; Flash, Exemplar
  - Alternative architecture with conventional technology
    - e.g., PIM, MTA
  - Alternative architecture and advanced technologies
    - hybrid technology multi-threaded
  - Special purpose devices
    - e.g., Grape-6
- Perform detailed applications studies at scale
  - compare against selected architectures

Recommendations (cont.)

- Develop petaflops scale latency management
  - Architecture
    - aggressive prefetching
    - data streaming
    - advanced cache coherence (e.g. COMA)
    - multi-threaded
  - Algorithmic methods
  - System software techniques for hiding latency
- Demonstrate that applications can expose 100,000-way parallelism and that architectures and system software can exploit 100K concurrent threads

Recommendations (cont.)

- Accelerate research in promising advanced *sidestream* technologies:
  - Superconductor RSFQ
  - Holographic optical storage
  - Optical guided and free-space interconnect
  - Exotic semiconductors
- Explore algorithms for special purpose and reconfigurable structures
- Initiate early software development of layered software architecture for scalability and code reuse