HTMT

A Hybrid Technology Approach to Petaflops Computing

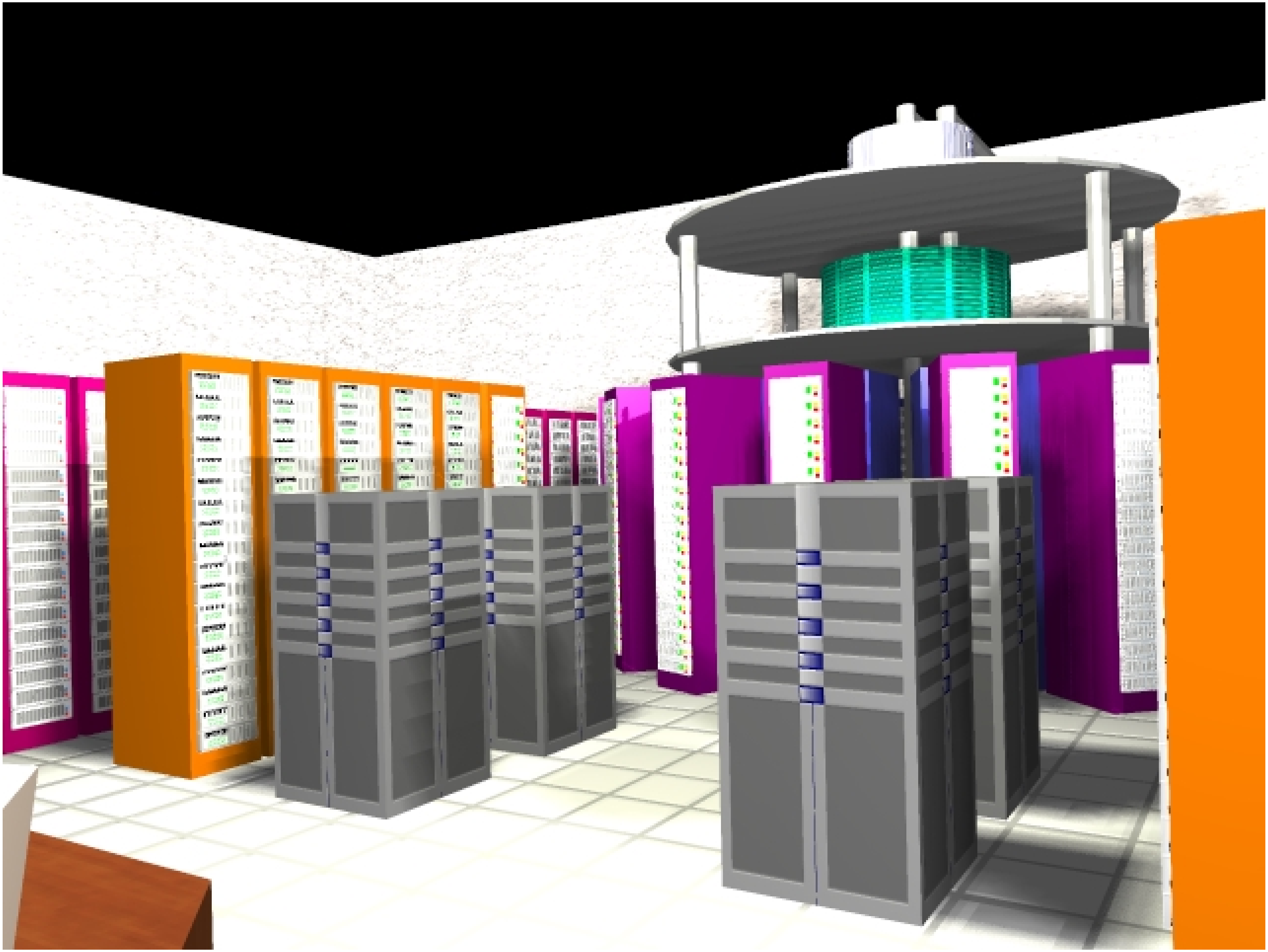
Thomas Sterling

Larry Bergman

NASA Jet Propulsion Laboratory

February 16, 1999

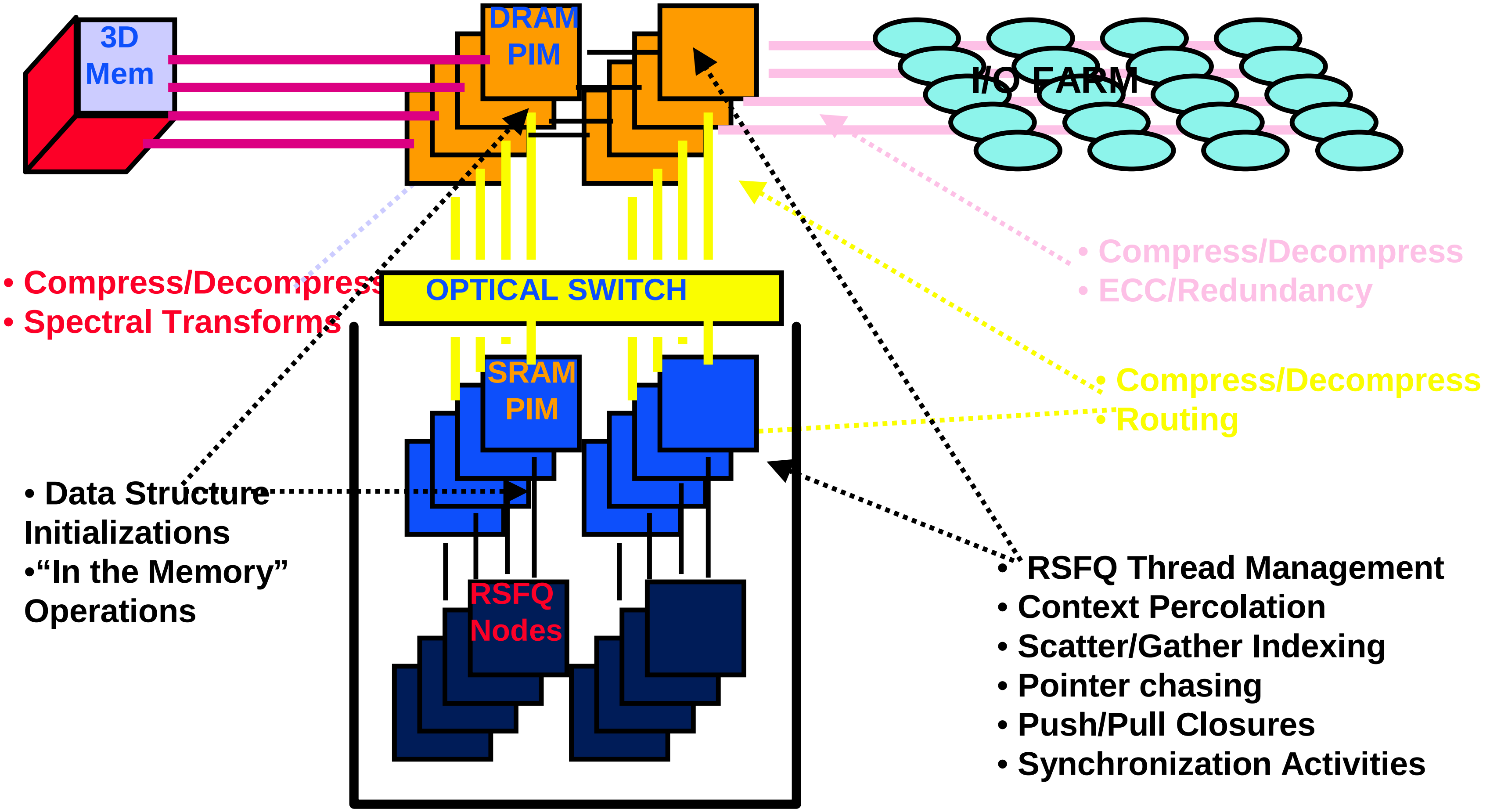
Warning: This is NOT a ***Beowulf*** *talk and may not be suitable for all audiences.*



HTMT Objectives

- Scalable architecture with high sustained performance in the presence of disparate cycle times and latencies
- Exploit diverse device technologies to achieve substantially superior operating point
- Execution model to simplify parallel system programming and expand generality and applicability

Hybrid Technology MultiThreaded Architecture



- Compress/Decompress
- Spectral Transforms

- Data Structure Initializations
- “In the Memory” Operations

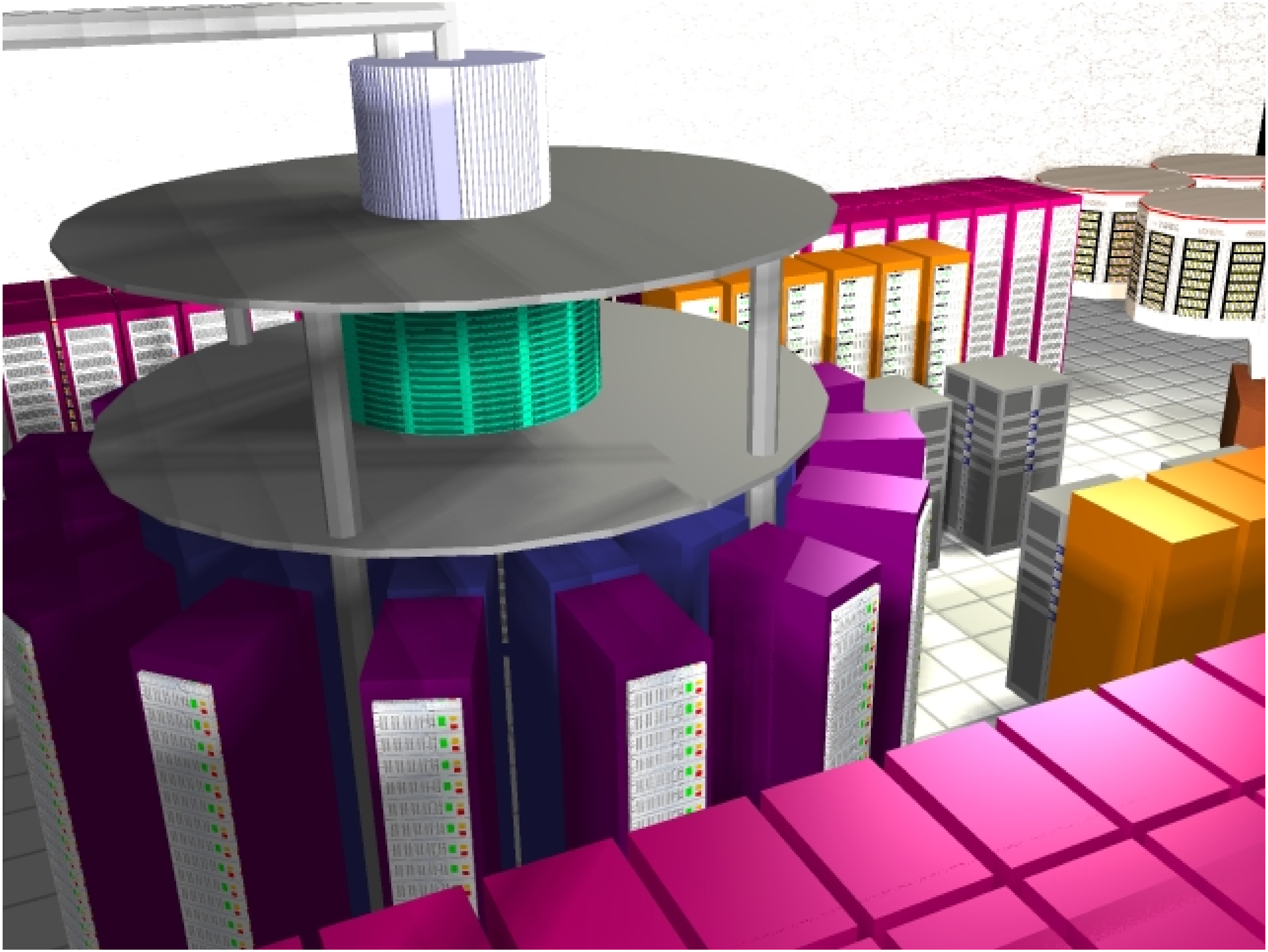
- Compress/Decompress
- ECC/Redundancy

- Compress/Decompress
- Routing

- RSFQ Thread Management
- Context Percolation
- Scatter/Gather Indexing
- Pointer chasing
- Push/Pull Closures
- Synchronization Activities

Summary of HTMT

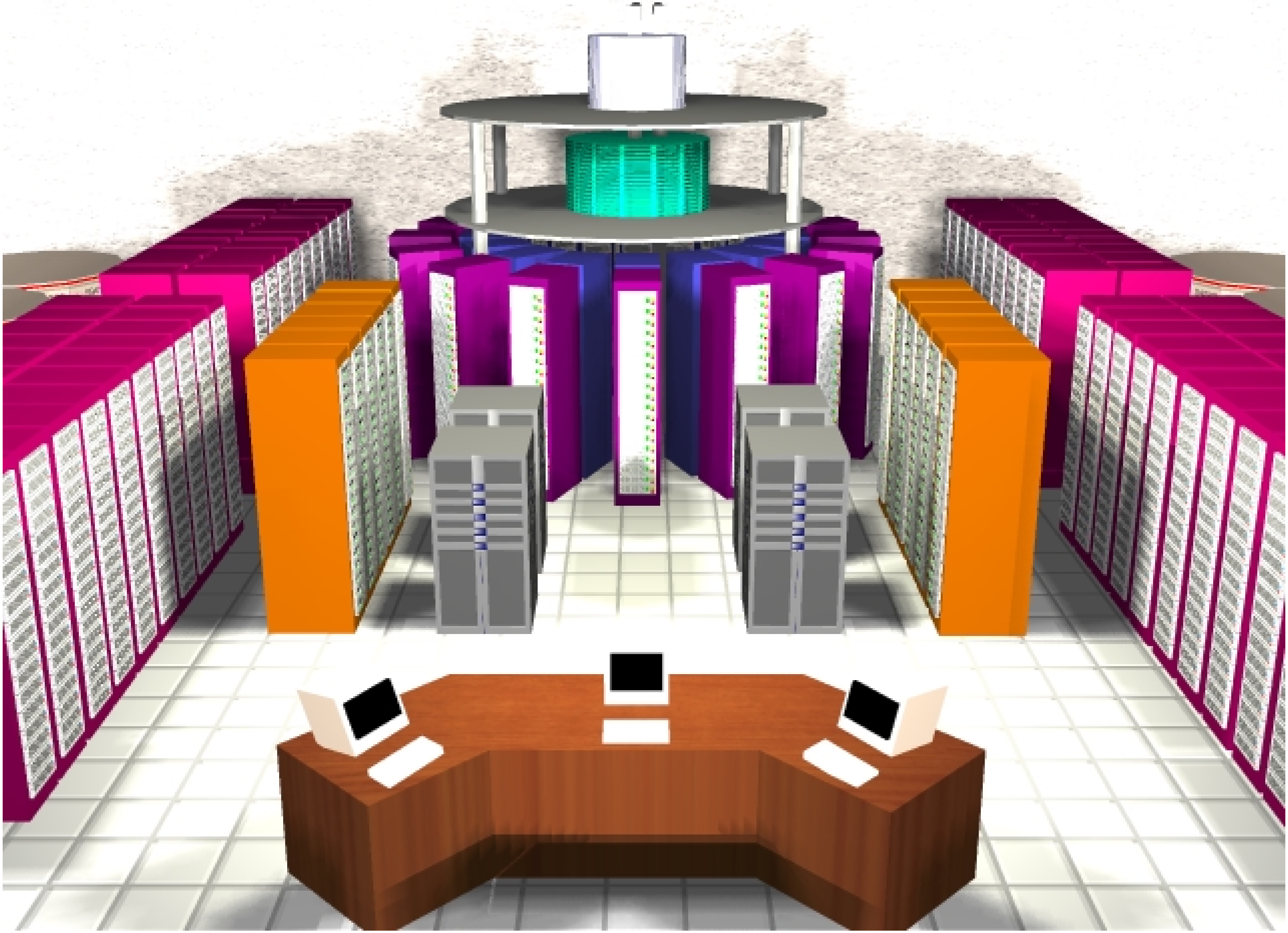
- processor: 150 GHz, 600 Gflops
- # processors: 2048
- memory: 16 Tbytes PIM-DRAM, 80ns access time
- interconnect: Data Vortex, 500 Gbps/channel, > 10 Pbps bi-section bw
- 3/2 storage: 1 Pbyte, 10 us access time
- shared memory, 4 level hierarchy
- latency management: multithreaded with percolation



Storage Capacity by Subsystem

2007 Design Point

Subsystem	Unit Storage	# of Units	Total Storage
CRAM	32 KB	16 K	512 MB
SRAM	64 MB	16 K	1 TB
DRAM	512 MB	32 K	16 TB
HRAM	10 GB	128 K	1 PB
Primary Disk	100 GB	100 K	10 PB
Secondary Disk	100 GB	100 K	10 PB
Tape	1 TB	6Kx20	120 PB



HTMT Strategy

- High performance
  - Superconductor RSFQ logic
  - *Data Vortex* optical interconnect network
  - PIM smart memory
- Low power
  - Superconductor RSFQ logic
  - Optical holographic storage
  - PIM smart memory

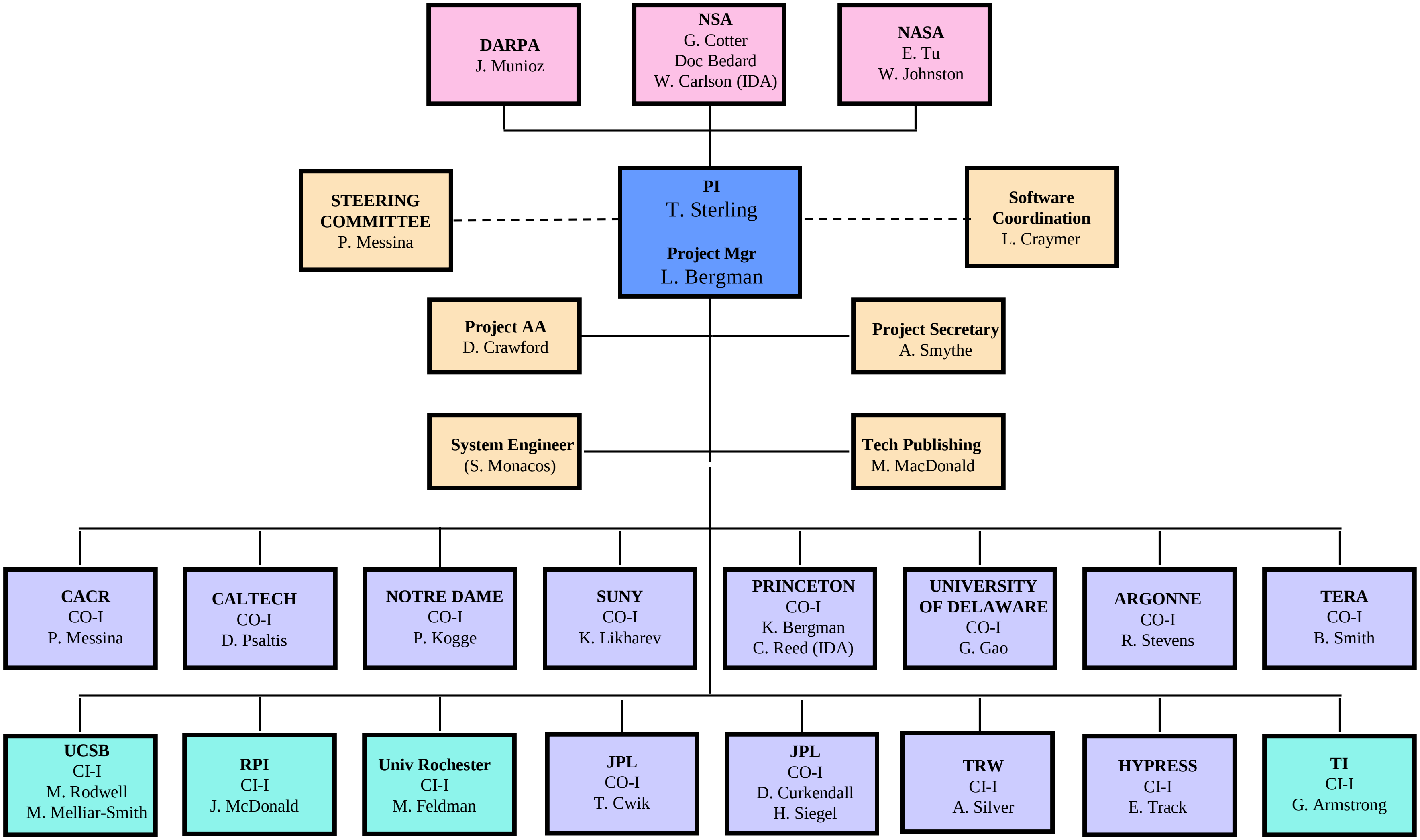
HTMT Strategy (cont)

- Low cost
  - reduce wire count through chip-to-chip fiber
  - reduce processor count through x100 clock speed
  - reduce memory chips by 3-2 holographic memory layer
- Efficiency
  - processor level multithreading
  - smart memory managed second stage *context pushing* multithreading
  - fine grain regular & irregular data parallelism exploited in memory
  - high memory bandwidth and low latency ops through PIM
  - memory to memory interactions without processor intervention
  - hardware mechanisms for synchronization, scheduling, data/context migration, gather/scatter

HTMT Strategy (cont)

- Programmability
  - Global shared name space
  - hierarchical parallel thread flow control model
    - no explicit processor naming
  - automatic latency management
    - automatic processor load balancing
    - runtime fine grain multithreading
    - automatic context pushing for process migration (percolation)
  - configuration transparent, runtime scalable

HTMT Organization



DARPA
J. Munioz

NSA
G. Cotter
Doc Bedard
W. Carlson (IDA)

NASA
E. Tu
W. Johnston

STEERING COMMITTEE
P. Messina

PI
T. Sterling

Project Mgr
L. Bergman

Software Coordination
L. Craymer

Project AA
D. Crawford

Project Secretary
A. Smythe

System Engineer
(S. Monacos)

Tech Publishing
M. MacDonald

CACR
CO-I
P. Messina

CALTECH
CO-I
D. Psaltis

NOTRE DAME
CO-I
P. Kogge

SUNY
CO-I
K. Likharev

PRINCETON
CO-I
K. Bergman
C. Reed (IDA)

UNIVERSITY OF DELAWARE
CO-I
G. Gao

ARGONNE
CO-I
R. Stevens

TERA
CO-I
B. Smith

UCSB
CI-I
M. Rodwell
M. Melliar-Smith

RPI
CI-I
J. McDonald

Univ Rochester
CI-I
M. Feldman

JPL
CO-I
T. Cwik

JPL
CO-I
D. Curkendall
H. Siegel

TRW
CI-I
A. Silver

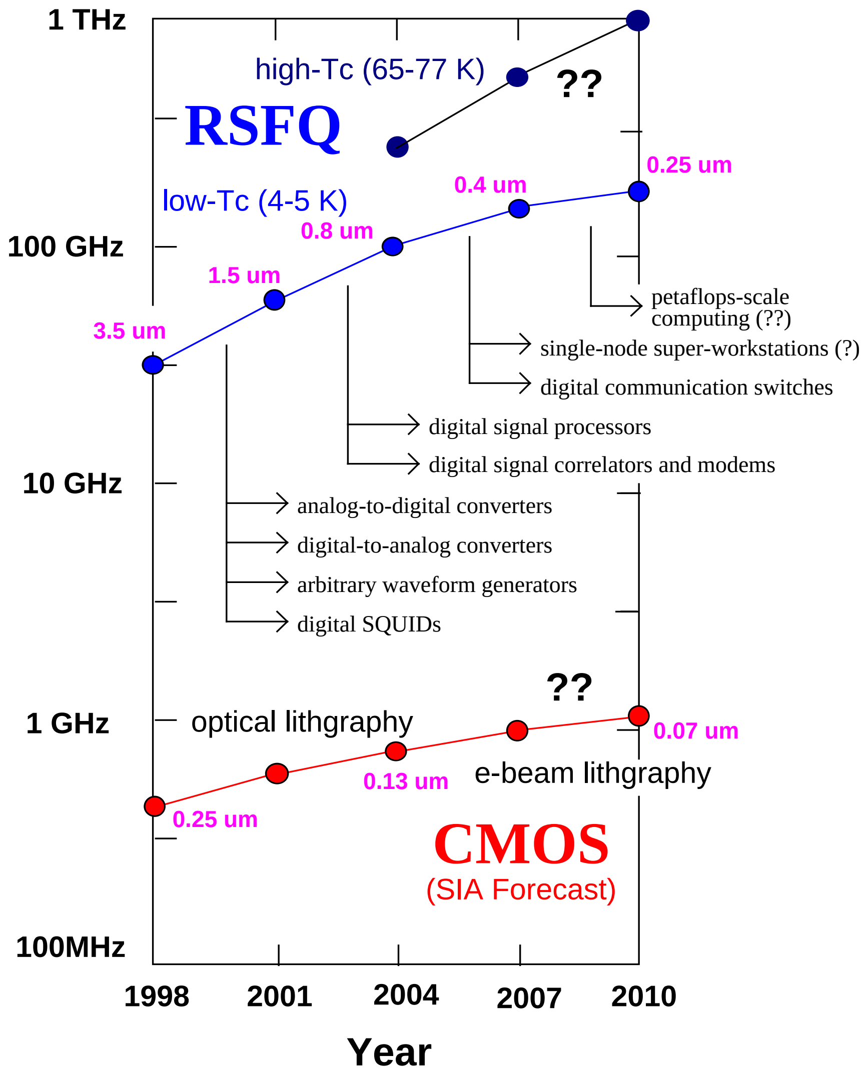
HYPRESS
CI-I
E. Track

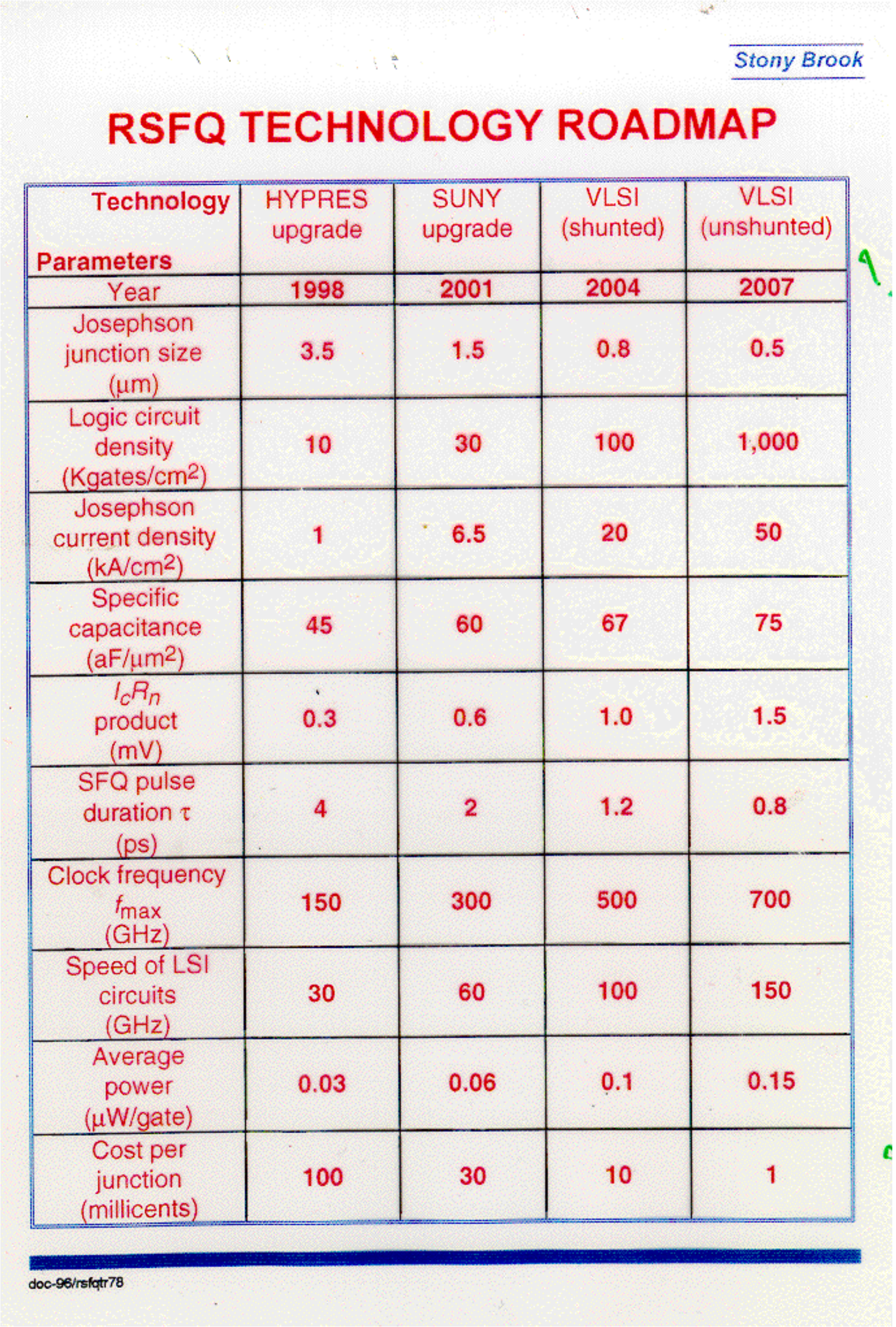
TI
CI-I
G. Armstrong

Areas of Accomplishments

- Concepts and Structures
  - approach strategy
  - device technologies
  - subsystem design
  - efficiency, productivity, generality
- System Architecture
  - size, cost, complexity, power
- System Software
  - resource management
  - multiprocessor emulator
- Applications
  - multithreaded codes
  - scaling models
- Evaluation
  - feasibility
  - cost
  - performance
- Future Directions
  - Phase 3 prototype
  - Phase 4 petaflops system
  - Proposals

RSFQ Roadmap (VLSI Circuit Clock Frequency)





Stony Brook

RSFQ TECHNOLOGY ROADMAP

Technology Parameters	HYPRES upgrade	SUNY upgrade	VLSI (shunted)	VLSI (unshunted)
Year	**1998**	**2001**	**2004**	**2007**
Josephson junction size (µm)	**3.5**	**1.5**	**0.8**	**0.5**
Logic circuit density (Kgates/cm^2)	**10**	**30**	**100**	**1,000**
Josephson current density (kA/cm^2)	**1**	**6.5**	**20**	**50**
Specific capacitance (aF/µm^2)	**45**	**60**	**67**	**75**
I_cR_n product (mV)	**0.3**	**0.6**	**1.0**	**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**

doc-96/rsfqtr78

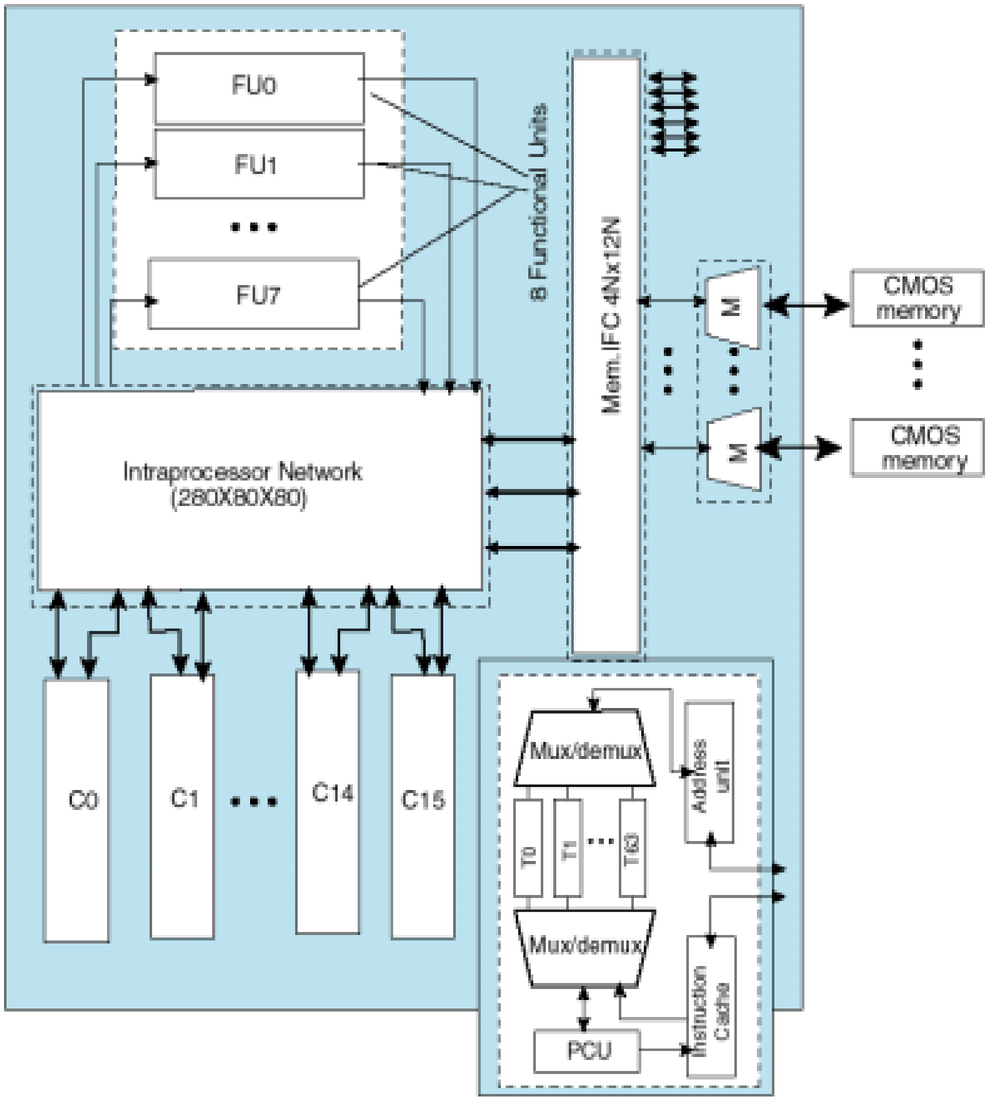
Advantages

- X100 clock speeds achievable
- X100 power efficiency advantage
- Easier fabrication
- Leverage semiconductor fabrication tools
- First technology to encounter ultra-high speed operation

Superconductor Processor

- 100 GHz clock, 33 GHz inter-chip
- 0.8 micron Niobium on Silicon
- 100K gates per chip
- 0.05 watts per processor
- 100Kwatts per Petaflops

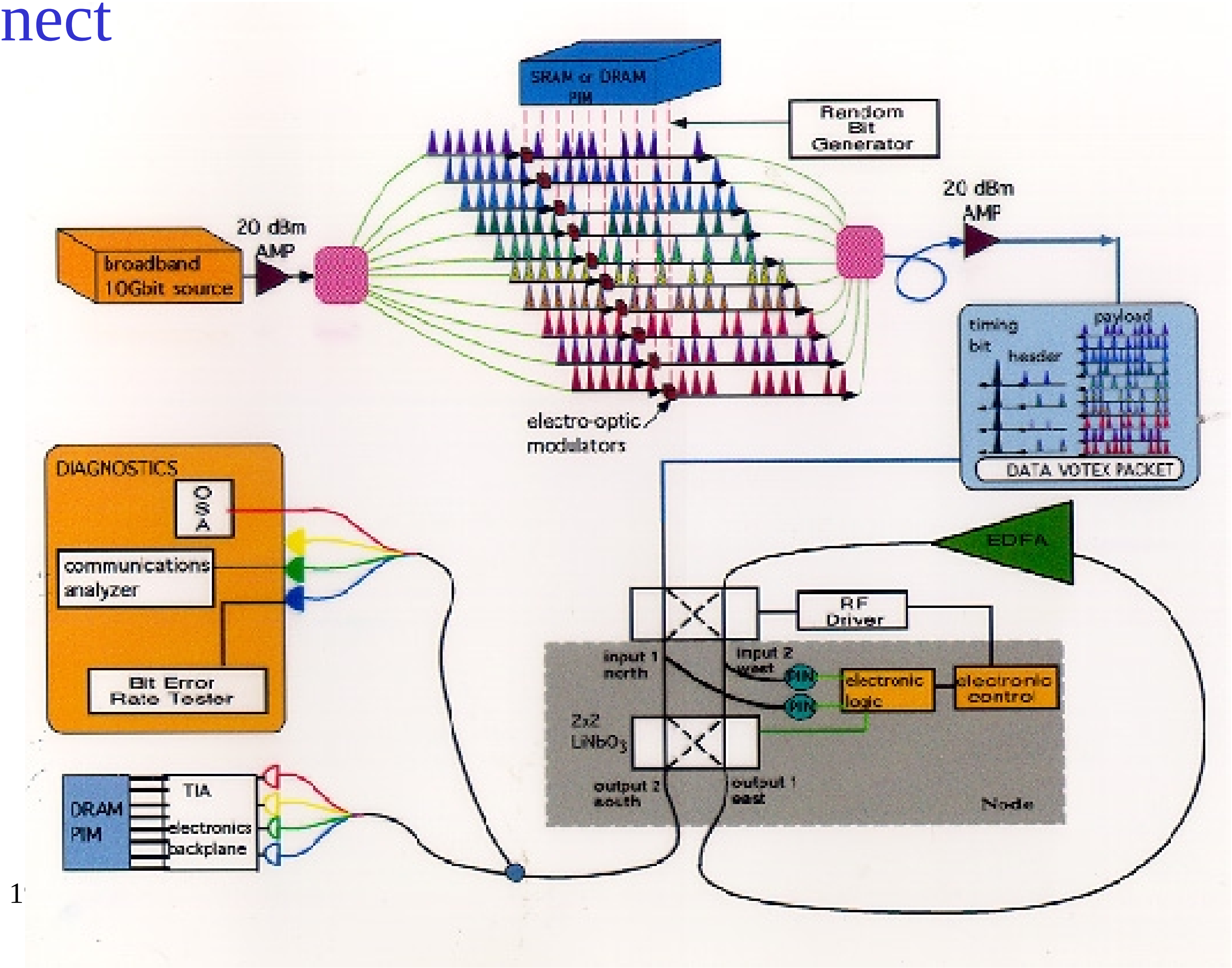
Single SCOPE Structure

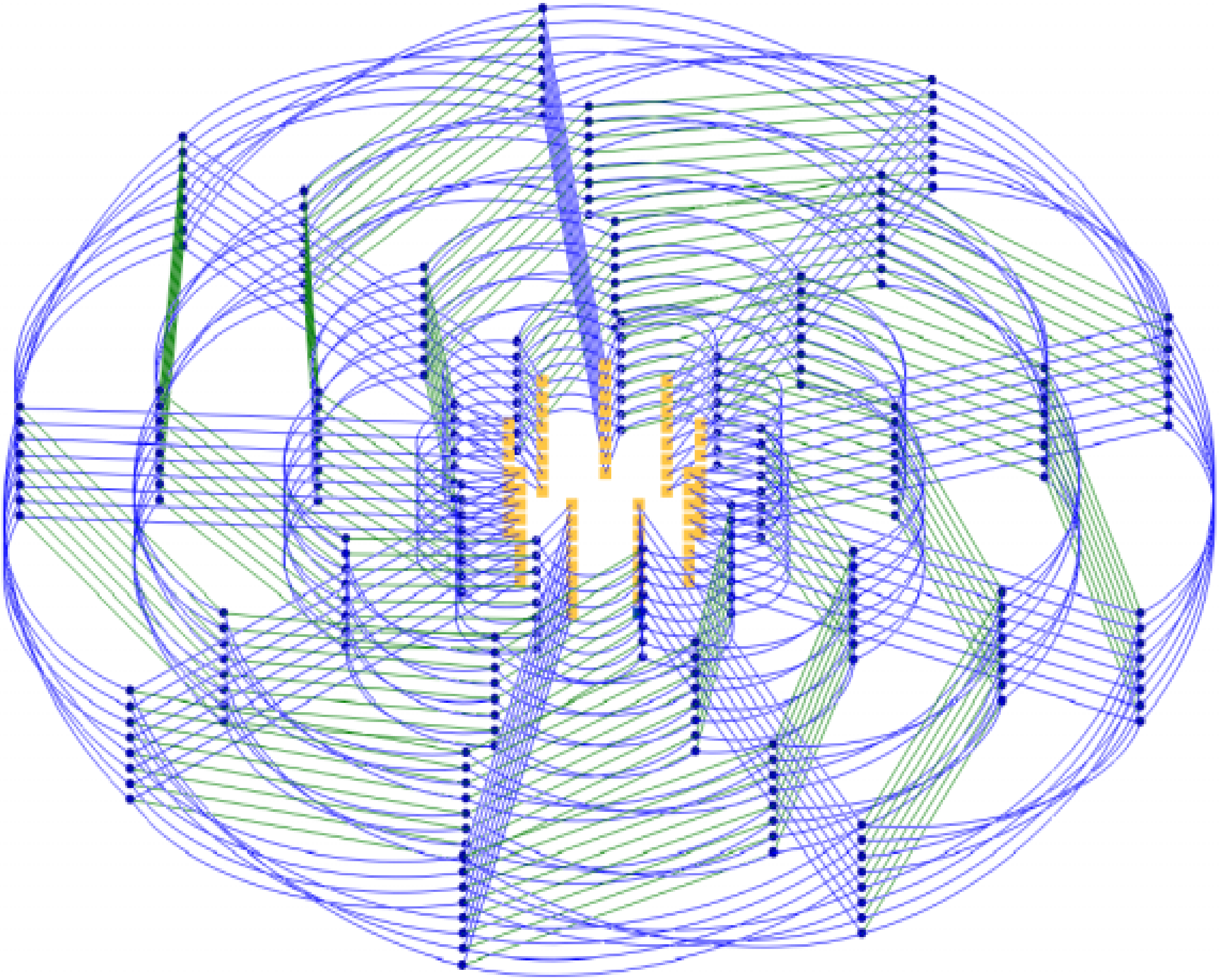


Accomplishments - Processor

- SPELL Architecture
- Detailed circuit design for critical paths
- CRAM Memory design initiated
- 1st network design and analysis/simulation
- 750 GHz logic demonstrated
- Detailed sizing, cost, and power analysis
- Estimate for fabrication facilities investment
- Barriers and path to 0.4-0.25 micron regime
- Sizing for Phase 3 50 Gflops processor

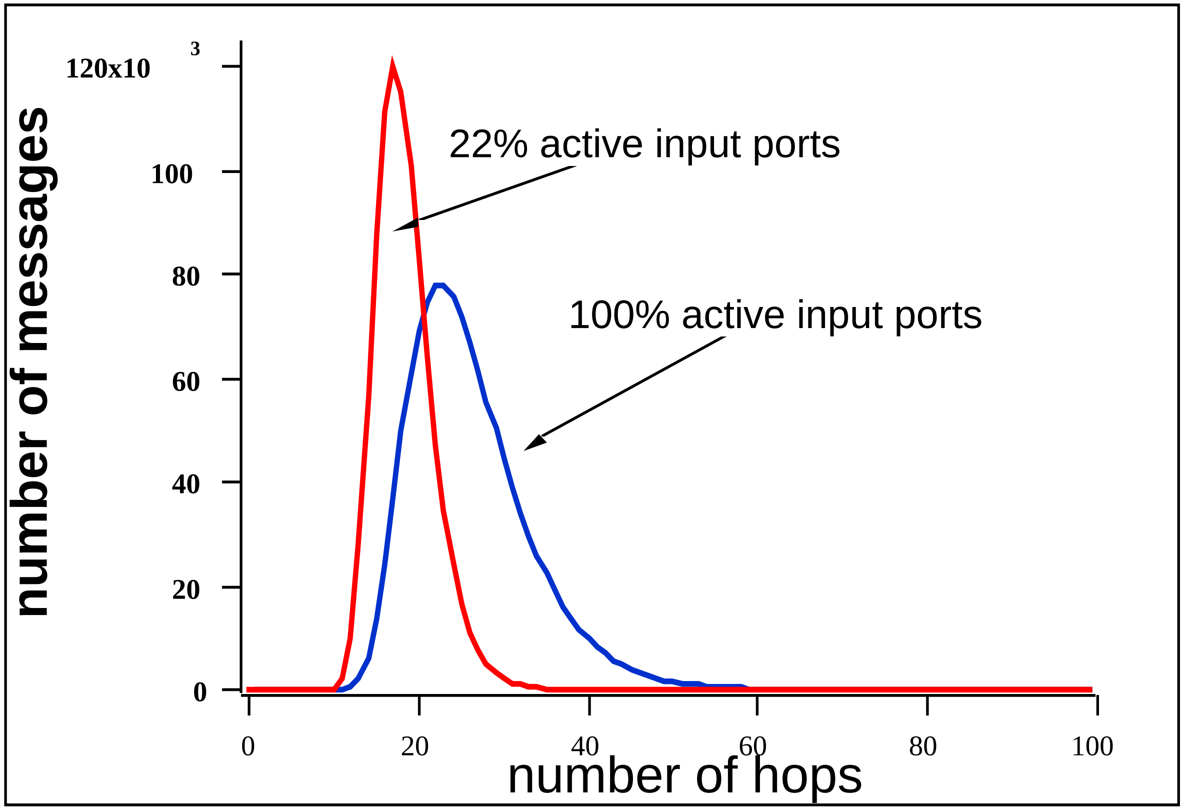
Data Vortex Optical Interconnect

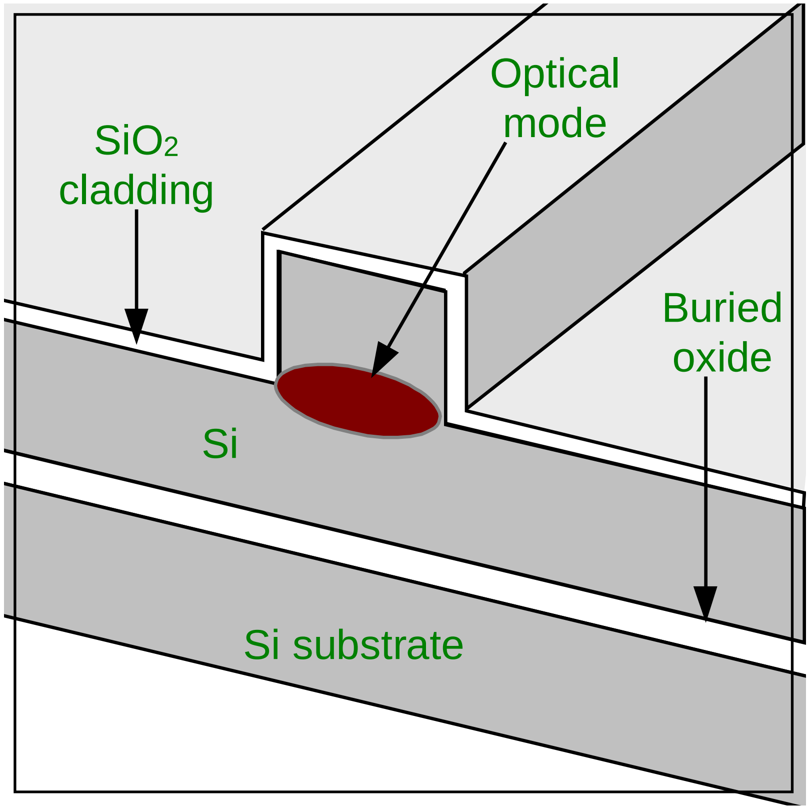




DATA VORTEX LATENCY DISTRIBUTION

network height = 1024





Single-mode rib waveguides on silicon-on-insulator wafers‡

Hybrid sources and detectors

Mix of CMOS-like and ‘micromachining’-type processes for fabrication

‡ **e.g:**

R A Soref, J Schmidtchen & K Petermann, IEEE J. Quantum Electron. 27 p1971 (1991)

A Rickman, G T Reed, B L Weiss & F Navamar, IEEE Photonics Technol. Lett. 4 p.633 (1992)

B Jalali, P D Trinh, S Yegnanarayanan & F Coppinger IEE Proc. Optoelectron. 143 p.307 (1996)

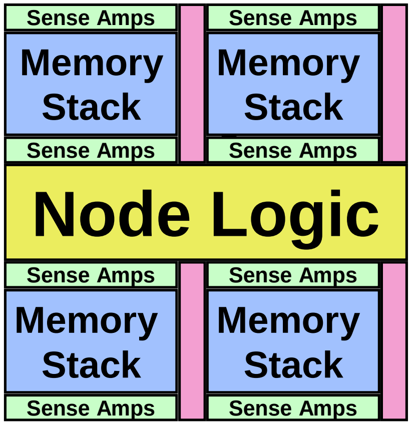
Data Vortex Parameters for Petaflops in 2007

- Bi-section sustained bandwidth: 4000 Tbps
- Per port data rate: 640 Gbps
- Single wavelength channel rate: 10 Gbps
- Level of WDM: 64 colors
- Number of input ports: 6250
- Angle nodes: 7
- Network node height: 4096
- Number of nodes per cylinder: 28672
- Number of cylinders: 13
- Total node number: 372736

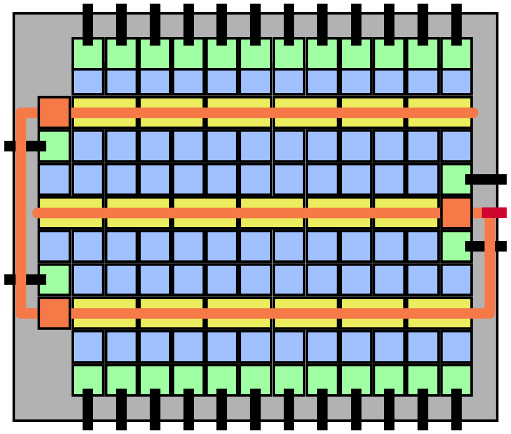
Accomplishments - Data Vortex

- Implemented and tested optical device technology
- Prototyped electro-optical butterfly switch
- Design study of electro-optic integrated switch
- Implemented and tested most of end-to-end path
- Design of topology to size
- Simulation of network behavior under load
- Modified structure for ease of packaging
- Size, complexity, power studies
- Initial interface design

PIM Provides Smart Memory



Basic Silicon Macro



Single Chip

- Merge logic and memory
- Integrate multiple logic/mem stacks on single chip
- Exposes high intrinsic memory bandwidth
- Reduction of memory access latency
- Low overhead for memory oriented operations
- Manages data structure manipulation, context coordination and percolation

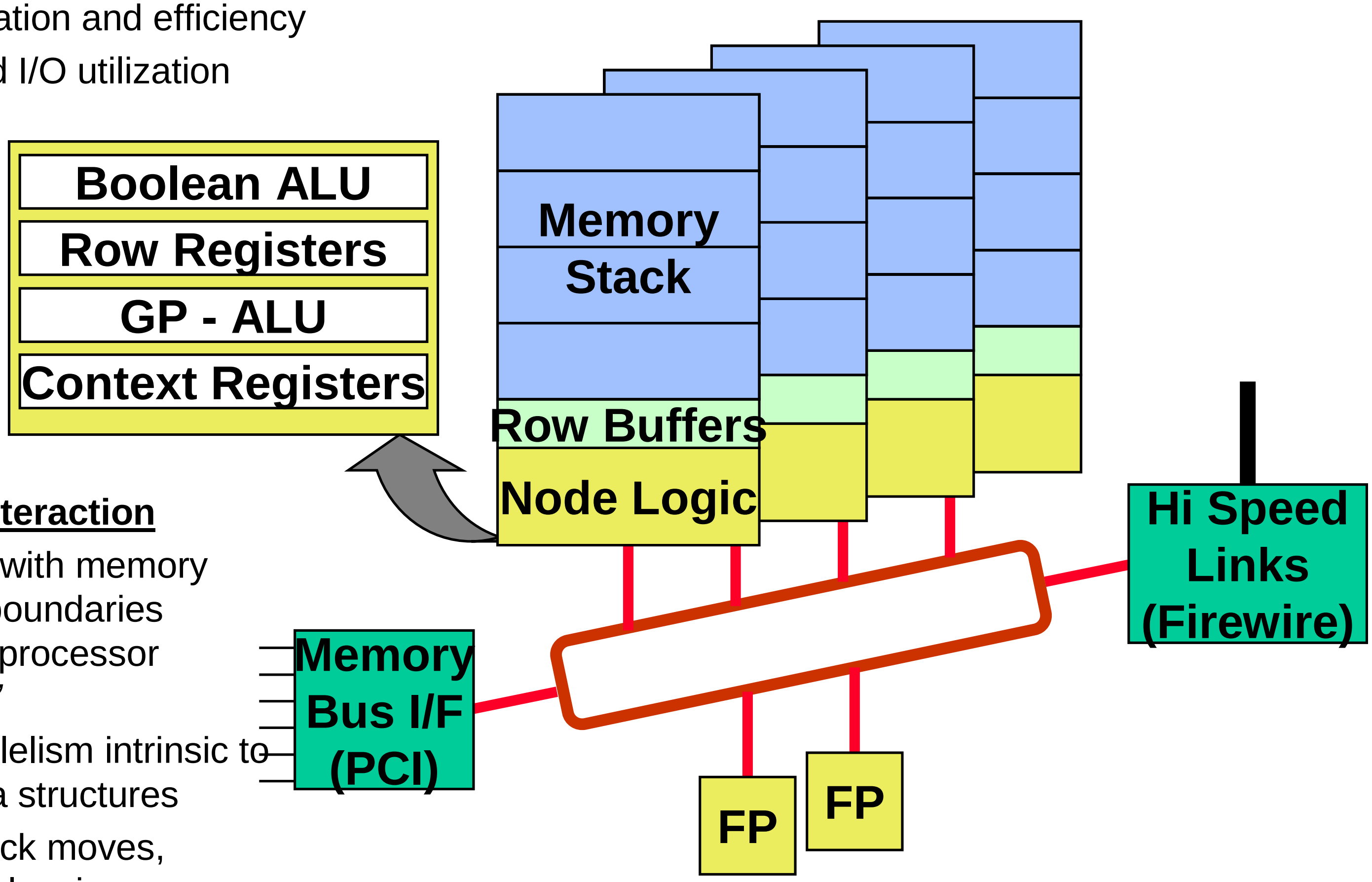
Multithreaded PIM DRAM

Multithreaded Control of PIM Functions

- multiple operation sequences with low context switching overhead
- maximize memory utilization and efficiency
- maximize processor and I/O utilization
- multiple banks of row buffers to hold data, instructions, and addr
- data parallel basic operations at row buffer
- manages shared resources such as FP

Direct PIM to PIM Interaction

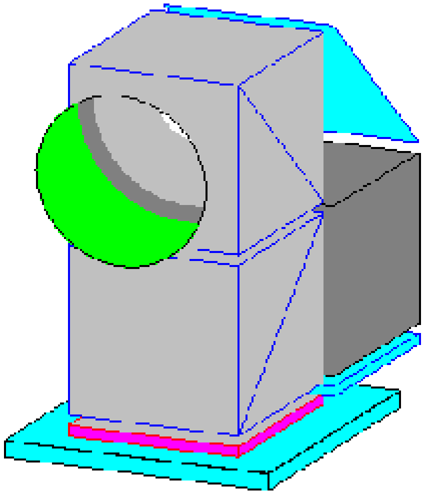
- memory communicates with memory within and across chip boundaries without external control processor intervention by *"parcels"*
- exposes fine grain parallelism intrinsic to vector and irregular data structures
- e.g. pointer chasing, block moves, synchronization, data balancing

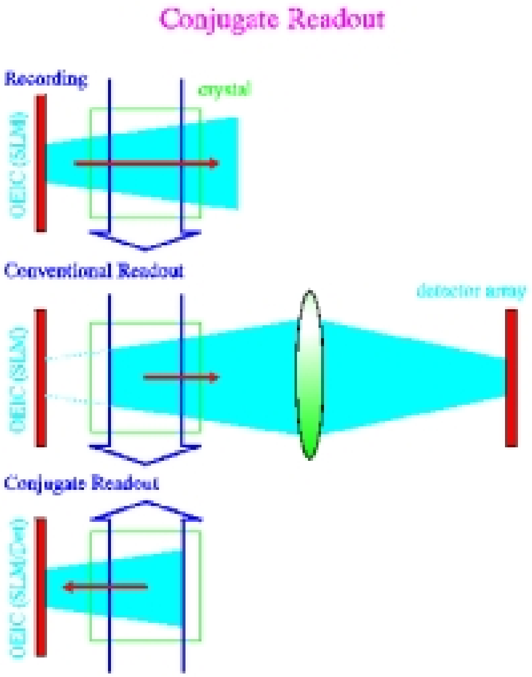


Accomplishments - PIM DRAM

- Establish operational opportunity and requirements
- Win $12.2M DARPA contract for DIVA
  - USC ISI prime
  - Caltech, Notre Dame, U. of Delaware
  - Deliver 8 Mbyte part in FY01 at 0.25 micron
- Architecture concept design complete
  - *parcel* message driven computation
  - multithreaded resource management
- Analysis of size, power, bandwidth
- Diva to be used directly in Phase 3 testbed

Holographic 3/2 Memory





Performance Scaling

	1998	2001	2004
Module capacity	1 Gbit	1 GB	10 GB
Number of modules		10^5	10^5
Access time	1 ms	100 µs	10 µs
Readout bandwidth	1 Gb/s	.1 PB/s	1 PB/s
Record bandwidth	1 Mb/s	1 GB/s	.1 PB/s

Advantages

- petabyte memory
- competitive cost
- 10 µsec access time
- low power
- efficient interface to DRAM

Disadvantages

- recording rate is slower than the readout rate for $LiNbO_3$
- recording must be done in GB chunks
- long term trend favors DRAM unless new materials and lasers are used

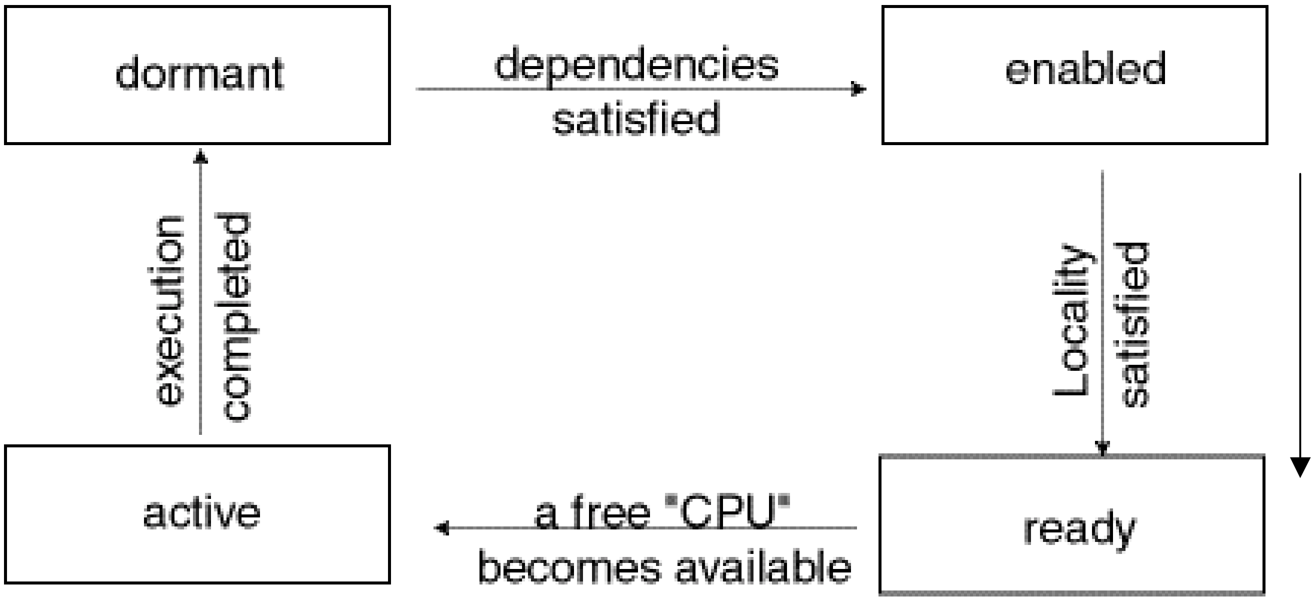
Accomplishments - HoloStore

- Detailed study of two optical storage technologies
  - photo refractive
  - spectral hole burning
- Operational photo refractive read/write storage
- Access approaches explored for 10 usec regime
  - pixel array
  - wavelength multiplexing
- Packaging studies
- power, size, cost analysis

Multilevel Multithreaded Execution Model

- Extend latency hiding of multithreading
- Hierarchy of logical thread
  - 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

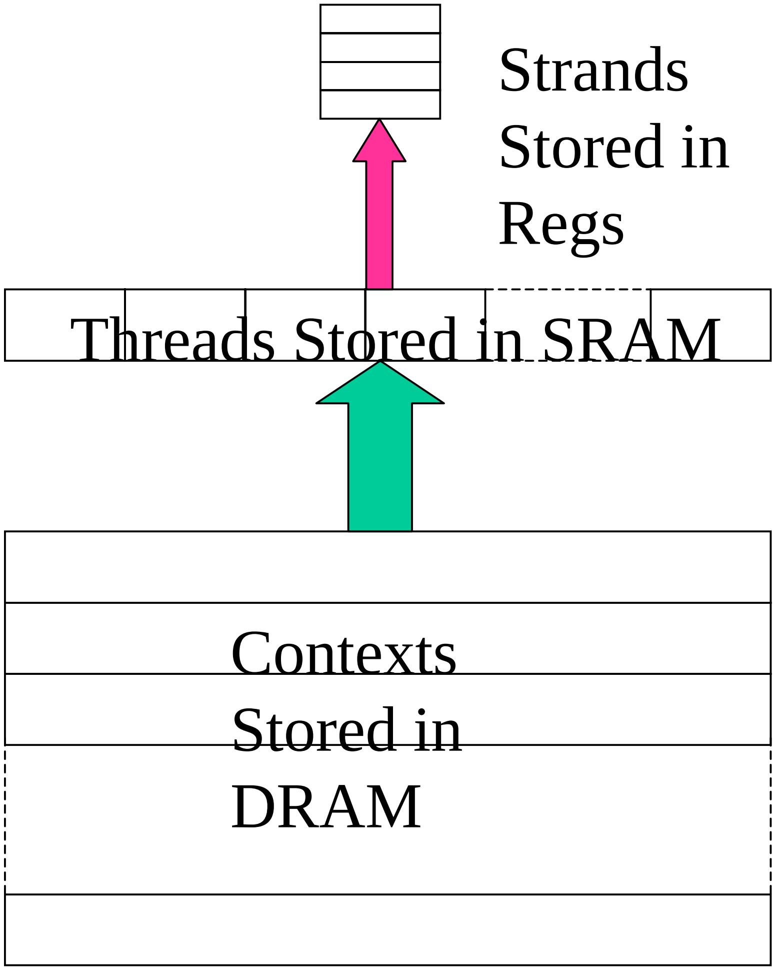
HTMT Thread Activation State Diagram



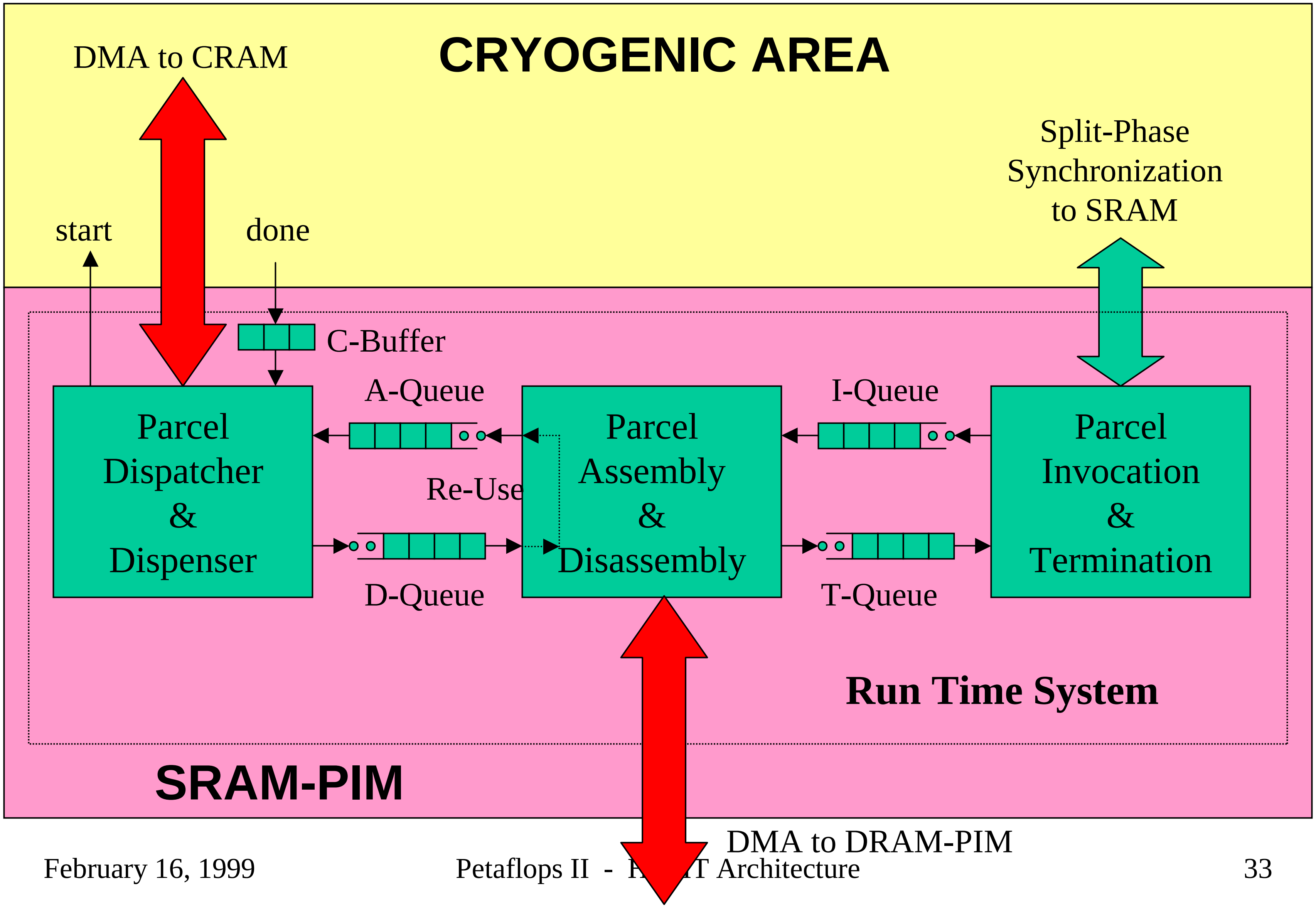
Percolation of threads

Percolation of Active Tasks

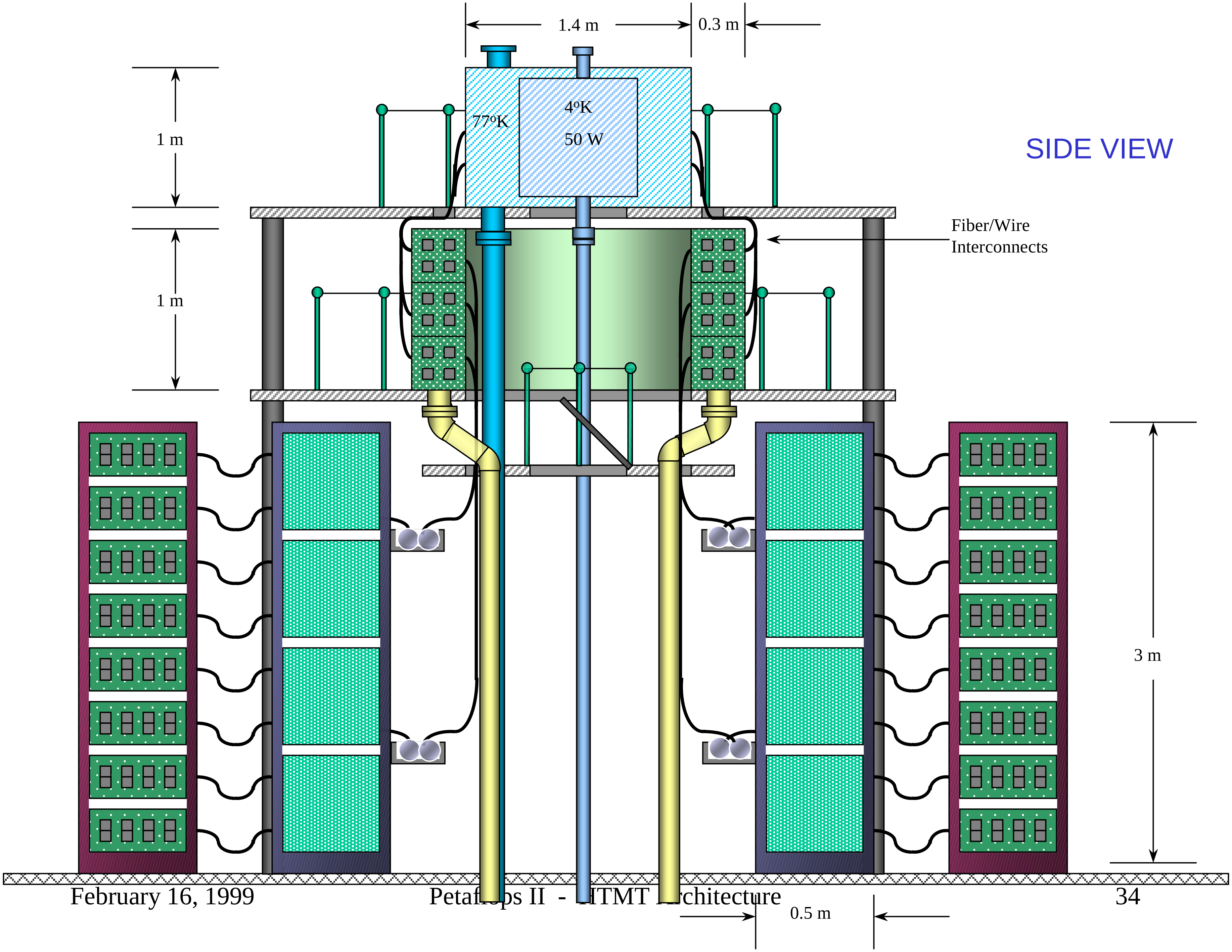
- Multiple stage latency management methodology
- Augmented multithreaded resource scheduling
- Hierarchy of task contexts
- Coarse-grain contexts coordinate in PIM memory
- Ready contexts migrate to SRAM under PIM control releasing threads for scheduling
- Threads pushed into SRAM/CRAM frame buffers
- Strands loaded in register banks on space available basis



HTMT Percolation Model

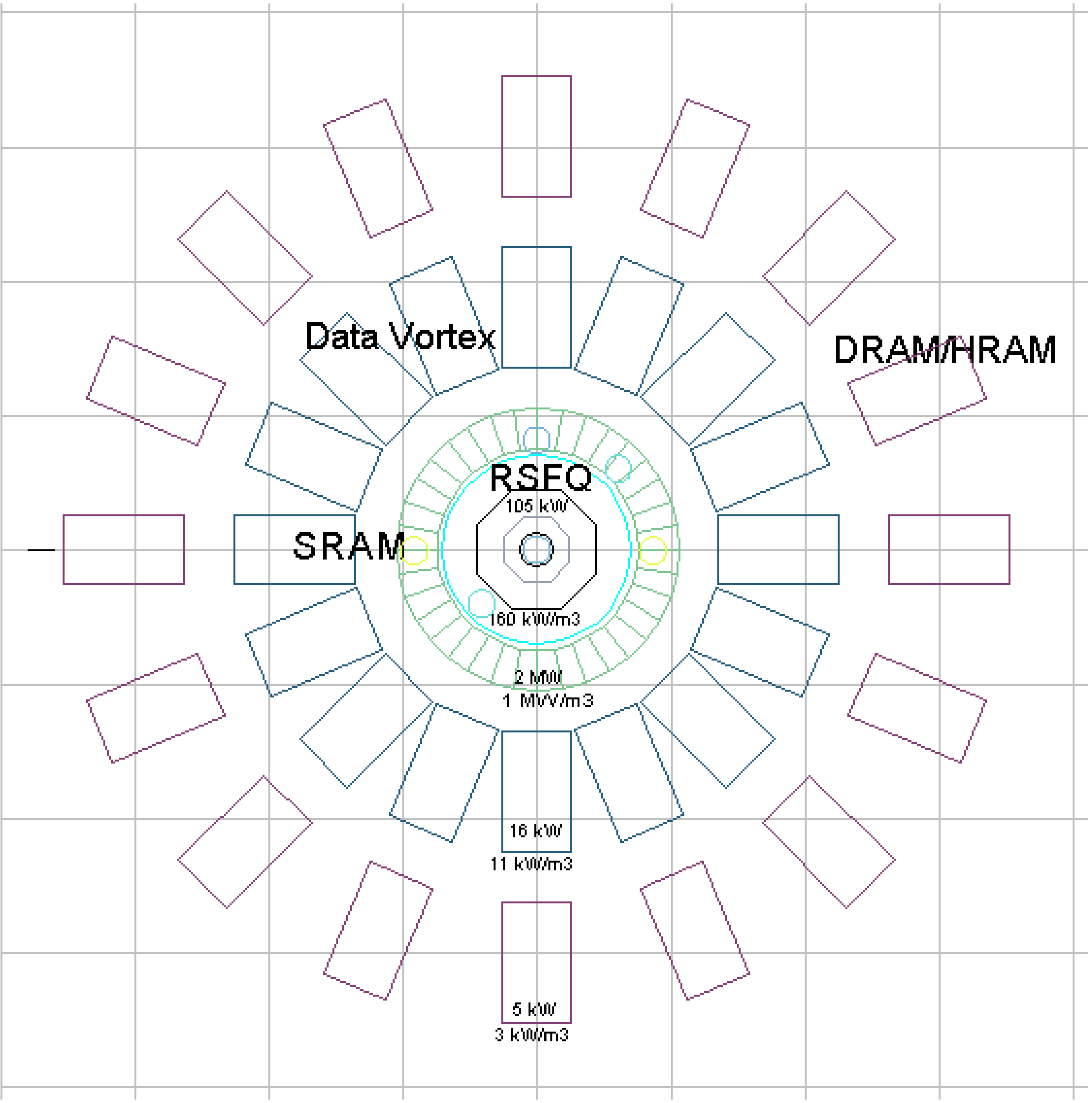

CRYOGENIC AREA
DMA to CRAM
start
done
Split-Phase Synchronization to SRAM
C-Buffer
A-Queue
I-Queue
Parcel Dispatcher & Dispenser
Re-Use
Parcel Assembly & Disassembly
Parcel Invocation & Termination
D-Queue
T-Queue
Run Time System
SRAM-PIM
DMA to DRAM-PIM

SIDE VIEW

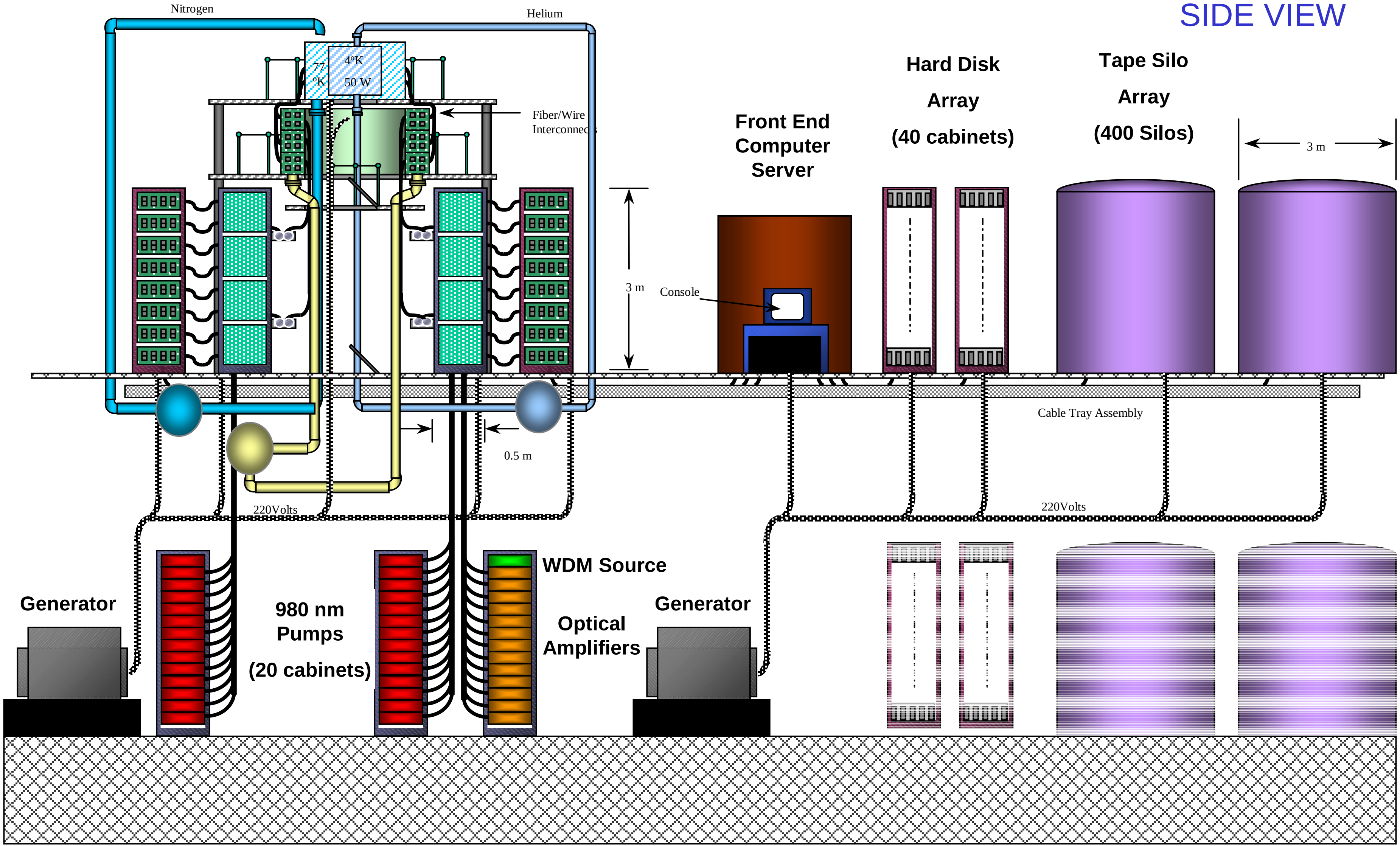


Top Down View of HTMT Machine

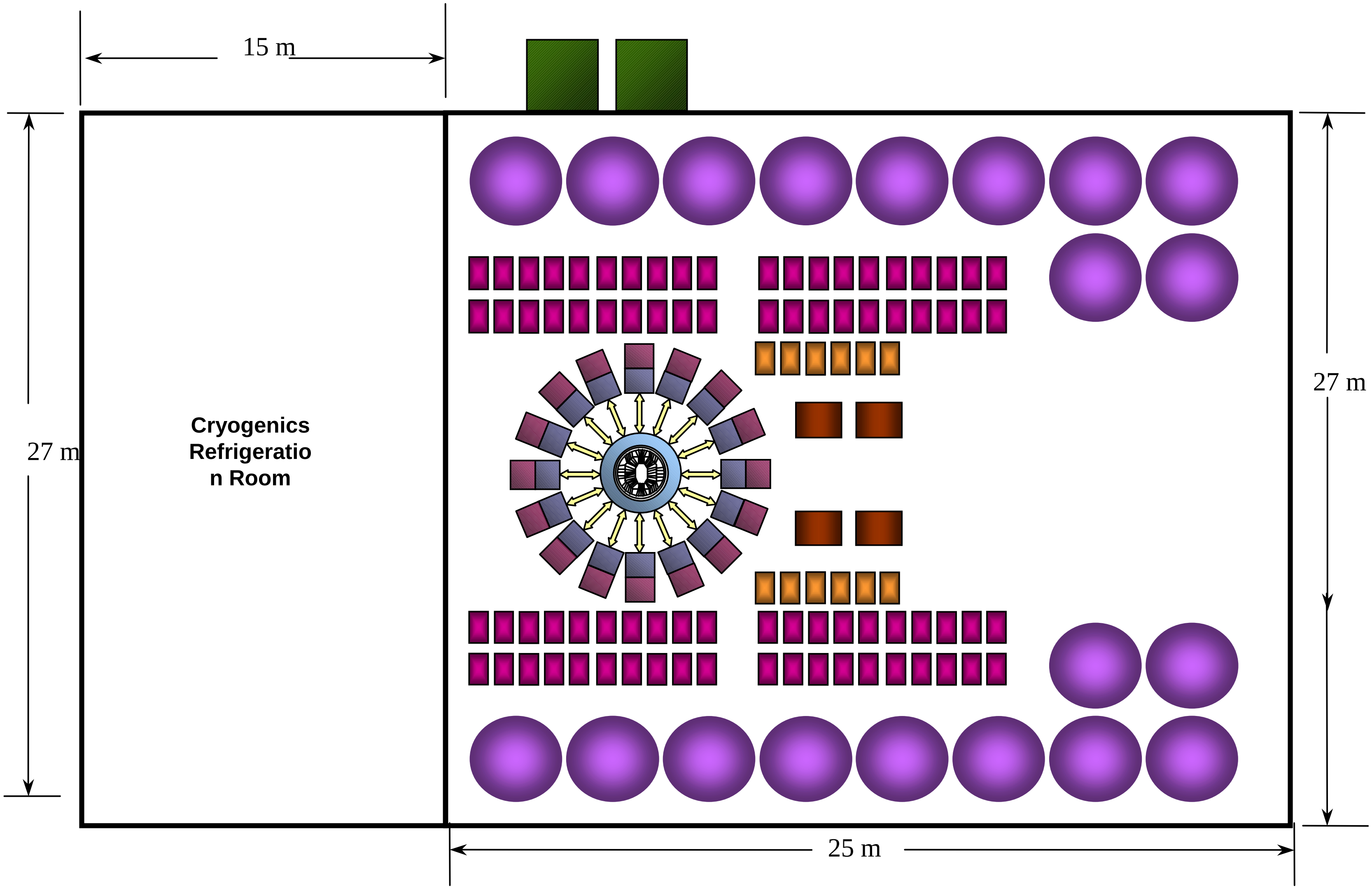
2007 Design Point



SIDE VIEW


Nitrogen
Helium
77 °K
4°K
50 W
Fiber/Wire Interconnects
3 m
Console
Front End Computer Server
Hard Disk Array (40 cabinets)
Tape Silo Array (400 Silos)
3 m
Cable Tray Assembly
0.5 m
220Volts
220Volts
Generator
980 nm Pumps (20 cabinets)
WDM Source
Optical Amplifiers
Generator

HTMT Facility (Top View)



Floor Area

1. HTMT	1,000
2. Server	250
3. Pump/MG	3,000
4. Laser 980	1,000
5. Disk Farm (80)	1,600
6. Tape Robot Farm (20)	4,000
7. Operator Room	1,000

TOTAL = 11,850 sq ft

Power Dissipation by Subsystem

Petaflops Design Point

Subsystem	Unit Type	Unit Power	# of Units	Total Power
Cryostat/Cooling	**System**	**400 kW**	**1**	**400 kW**
SRAM	**PIM**	**5 W**	**16 K**	**80 kW**
WDM source/amps	**Port**	**15 W**	**4 K**	**62 kW**
Data Vortex	**Subnet**	**2 kW**	**128**	**258 kW**
DRAM	**PIM**	**625 mW**	**32 K**	**20 kW**
HRAM	**HRAM**	**100 mW**	**128 K**	**13 kW**
Primary Disk	**Disk**	**15 W**	**100 K**	**1500 kW**
Tape	**Silo**	**1 kW**	**20**	**20 kW**
Server	**Machine**	**100 kW**	**1**	**100 kW**
			TOTAL	**2.4 MW**

Subsystem Interfaces

2007 Design Point

Subsystem	Interface to	Wires/Port	Speed/Wire (bps)	#ports	Aggregate BW (Byte/s)	Wire count	type of IF
RSFQ	SRAM	16000	20.0E+9	512	20.5E+15	8.2E+6	wire
SRAM	RSFQ	1000	2.0E+9	8000	2.0E+15	8.0E+6	TBD
SRAM	Data Vortex	1000	2.0E+9	8000	2.0E+15	8.0E+6	wire
Data Vortex	SRAM	1	640.0E+9	2048	163.8E+12	2.0E+3	fiber
Data Vortex	DRAM	1	640.0E+9	2048	163.8E+12	2.0E+3	fiber
DRAM	Data Vortex	1000	1.0E+9	33000	4.1E+15	33.0E+6	wire
DRAM	HRAM	1000	1.0E+9	33000	4.1E+15	33.0E+6	wire
DRAM	Server	1	800.0E+6	1000	100.0E+9	1.0E+3	wire
Server	DRAM	1	800.0E+6	1000	100.0E+9	1.0E+3	(fiber channel)
Server	Disk	1	800.0E+6	1000	100.0E+9	1.0E+3	(fiber channel)
Server	Tape	1	800.0E+6	200	20.0E+9	200.0E+0	(fiber channel)
HRAM	DRAM	800	100.0E+6	1.00E+05	1.0E+15	80.0E+6	wire

- Same colors indicate a connection between subsystems
- Horizontal lines group interfaces within a subsystem

Accomplishments - Systems

- System architecture completed
- Physical structure design
- Parts count, power, interconnect complexity analysis
- Infrastructure requirements and impact
- Feasibility assessment

Distributed Isomorphic Simulator

- Executable Specification
  - subsystem functional/operational description
  - inter-subsystem interface protocol definition
- Distributed Low-cost Cluster of processors
- Cluster partitioned and allocated to separate subsystems
- Subsystem development groups “own” cluster partitions, and develop functional specification
- Subsystem partitions interact by agreed-upon interface protocols
- Runtime percolation and thread scheduling system software put on top of emulation software.