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

- Compress/Decompress
- ECC/Redundancy
- Routing
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

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Unit Storage</th>
<th># of Units</th>
<th>Total Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRAM</td>
<td>32 KB</td>
<td>16 K</td>
<td>512 MB</td>
</tr>
<tr>
<td>SRAM</td>
<td>64 MB</td>
<td>16 K</td>
<td>1 TB</td>
</tr>
<tr>
<td>DRAM</td>
<td>512 MB</td>
<td>32 K</td>
<td>16 TB</td>
</tr>
<tr>
<td>HRAM</td>
<td>10 GB</td>
<td>128 K</td>
<td>1 PB</td>
</tr>
<tr>
<td>Primary Disk</td>
<td>100 GB</td>
<td>100 K</td>
<td>10 PB</td>
</tr>
<tr>
<td>Secondary Disk</td>
<td>100 GB</td>
<td>100 K</td>
<td>10 PB</td>
</tr>
<tr>
<td>Tape</td>
<td>1 TB</td>
<td>6Kx20</td>
<td>120 PB</td>
</tr>
</tbody>
</table>
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
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)
# RSFQ Technology Roadmap

<table>
<thead>
<tr>
<th>Technology Parameters</th>
<th>HYPRES upgrade</th>
<th>SUNY upgrade</th>
<th>VLSI (shunted)</th>
<th>VLSI (unshunted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1998</td>
<td>2001</td>
<td>2004</td>
<td>2007</td>
</tr>
<tr>
<td>Josephson junction size (μm)</td>
<td>3.5</td>
<td>1.5</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Logic circuit density (Kgates/cm²)</td>
<td>10</td>
<td>30</td>
<td>100</td>
<td>1,000</td>
</tr>
<tr>
<td>Josephson current density (kA/cm²)</td>
<td>1</td>
<td>6.5</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Specific capacitance (aF/μm²)</td>
<td>45</td>
<td>60</td>
<td>67</td>
<td>75</td>
</tr>
<tr>
<td>$I_cP_n$ product (mV)</td>
<td>0.3</td>
<td>0.6</td>
<td>1.0</td>
<td>1.5</td>
</tr>
<tr>
<td>SFQ pulse duration $\tau$ (ps)</td>
<td>4</td>
<td>2</td>
<td>1.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Clock frequency $f_{\text{max}}$ (GHz)</td>
<td>150</td>
<td>300</td>
<td>500</td>
<td>700</td>
</tr>
<tr>
<td>Speed of LSI circuits (GHz)</td>
<td>30</td>
<td>60</td>
<td>100</td>
<td>150</td>
</tr>
<tr>
<td>Average power (μW/gate)</td>
<td>0.03</td>
<td>0.06</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>Cost per junction (millicents)</td>
<td>100</td>
<td>30</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>
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
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

- 22% active input ports
- 100% active input ports
Si substrate

SiO₂ cladding

Optical mode

Buried oxide

Si

Si substrate

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,

A Rickman, G T Reed, B L Weiss & F Navamar,

B Jalali, P D Trinh, S Yegnanarayanan & F Coppinger
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

- 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 addresses
- 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

February 16, 1999  
Petaflops II - HTMT Architecture
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

<table>
<thead>
<tr>
<th>Module capacity</th>
<th>1998</th>
<th>2001</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access time</td>
<td>1 ms</td>
<td>100 μs</td>
<td>10 μs</td>
</tr>
<tr>
<td>Readout bandwidth</td>
<td>1 Gb/s</td>
<td>.1 PB/s</td>
<td>1 PB/s</td>
</tr>
<tr>
<td>Record bandwidth</td>
<td>1 Mb/s</td>
<td>1 GB/s</td>
<td>.1 PB/s</td>
</tr>
<tr>
<td>Number of modules</td>
<td>$10^5$</td>
<td>$10^5$</td>
<td></td>
</tr>
<tr>
<td>Readout bandwidth</td>
<td>1 Gb/s</td>
<td>.1 PB/s</td>
<td>1 PB/s</td>
</tr>
<tr>
<td>Record bandwidth</td>
<td>1 Mb/s</td>
<td>1 GB/s</td>
<td>.1 PB/s</td>
</tr>
</tbody>
</table>

**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

- Dormant
  - Dependencies satisfied
  - Execution completed
- Active
- Ready
  - Locality satisfied
  - A free "CPU" becomes available

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
Top Down View of HTMT Machine
2007 Design Point
# Floor Area

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. HTMT</td>
<td>1,000</td>
</tr>
<tr>
<td>2. Server</td>
<td>250</td>
</tr>
<tr>
<td>3. Pump/MG</td>
<td>3,000</td>
</tr>
<tr>
<td>4. Laser 980</td>
<td>1,000</td>
</tr>
<tr>
<td>5. Disk Farm (80)</td>
<td>1,600</td>
</tr>
<tr>
<td>6. Tape Robot Farm (20)</td>
<td>4,000</td>
</tr>
<tr>
<td>7. Operator Room</td>
<td>1,000</td>
</tr>
</tbody>
</table>

**TOTAL = 11,850 sq ft**
## Power Dissipation by Subsystem

### Petaflops Design Point

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Unit Type</th>
<th>Unit Power</th>
<th># of Units</th>
<th>Total Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryostat/Cooling</td>
<td>System</td>
<td>400 kW</td>
<td>1</td>
<td>400 kW</td>
</tr>
<tr>
<td>SRAM</td>
<td>PIM</td>
<td>5 W</td>
<td>16 K</td>
<td>80 kW</td>
</tr>
<tr>
<td>WDM source/amps</td>
<td>Port</td>
<td>15 W</td>
<td>4 K</td>
<td>62 kW</td>
</tr>
<tr>
<td>Data Vortex</td>
<td>Subnet</td>
<td>2 kW</td>
<td>128</td>
<td>258 kW</td>
</tr>
<tr>
<td>DRAM</td>
<td>PIM</td>
<td>625 mW</td>
<td>32 K</td>
<td>20 kW</td>
</tr>
<tr>
<td>HRAM</td>
<td>HRAM</td>
<td>100 mW</td>
<td>128 K</td>
<td>13 kW</td>
</tr>
<tr>
<td>Primary Disk</td>
<td>Disk</td>
<td>15 W</td>
<td>100 K</td>
<td>1500 kW</td>
</tr>
<tr>
<td>Tape</td>
<td>Silo</td>
<td>1 kW</td>
<td>20</td>
<td>20 kW</td>
</tr>
<tr>
<td>Server</td>
<td>Machine</td>
<td>100 kW</td>
<td>1</td>
<td>100 kW</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>TOTAL</strong></td>
</tr>
</tbody>
</table>
## Subsystem Interfaces

### 2007 Design Point

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

- 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.