

FPGA and Cell Processor Performance Optimization for Brain-State-in-a Box (BSB) Cognitive Computing



Dr. Richard Linderman

Senior Scientist

AFRL/IF

Air Force Research Laboratory

Dr. Qing Wu and Dr. Qinru Qiu

**Dept. of Electrical and Computer
Engineering**

Binghamton University

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE AUG 2007		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE FPGA and Cell Processor Performance Optimization for Brain-State-in-a Box (BSB) Cognitive Computing				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Research Laboratory AFRL/IF				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES Symposium on Multicore and New Processing Technologies August 13-14 2007, The original document contains color images.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 21	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			



BSB Recall Operation

$$\mathbf{X}(t+1) = -\mathbf{A}\mathbf{X}(t) + \mathbf{X}(0) + \alpha \mathbf{X}(t) + \lambda \mathbf{X}(t) + \gamma \mathbf{X}(0)$$

- $\mathbf{X}(t+1)$ and $\mathbf{X}(t)$ are N dimensional real vectors;
- $\mathbf{A}$ is an $N \times N$ connection matrix;
- α is a scalar constant feedback factor;
- λ is an inhibition decay constant;
- γ is a nonzero constant if there is a need to maintain the input stimulation;
- $\mathbf{X}(0)$ is the input stimulation;
- $S()$ is the “squash” function

$$\mathbf{X}(t+1) = -\mathbf{A}\mathbf{X}(t) + \mathbf{X}(0) + \alpha \mathbf{X}(t) + \lambda \mathbf{X}(t) + \gamma \mathbf{X}(0)$$



BSB Training Operation

In Training, the $N \times N$ connection matrix A is modified as

$$\Delta A = lr * (X - AX) \otimes X$$

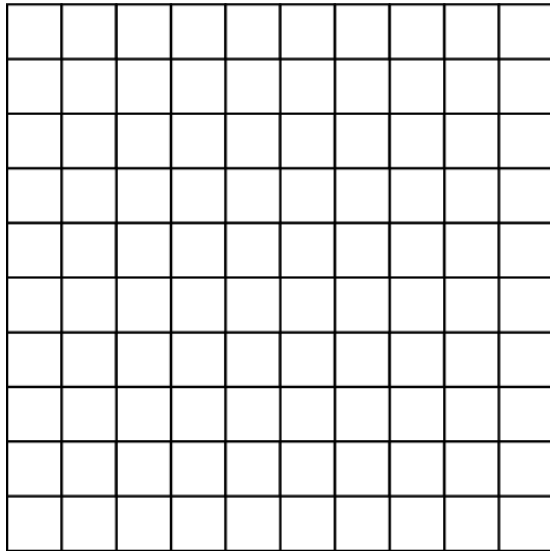
$$A = A + \Delta A$$

- X is the normalized input training pattern;
- lr is the Learning rate;
- $\otimes$ is the outer product of two vectors;



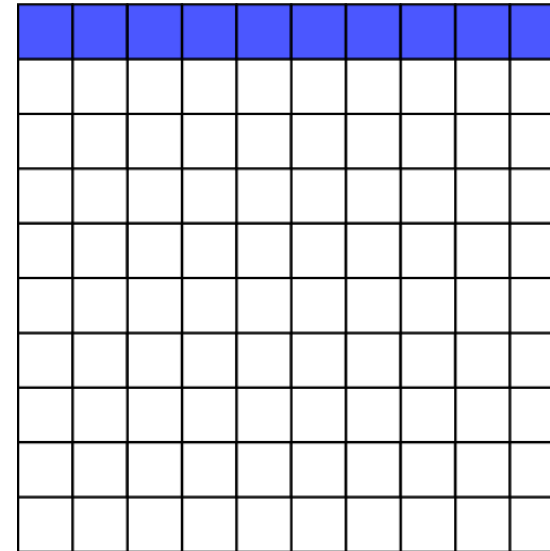
Application

Pattern recognition



10x10 black & white patterns

Example : Line Pattern "H1"



Total of 20 line patterns → 20 tag entries

Input/state vector structure

$X_{127} \sim X_{100}$:

Tags
 $X_{99} \sim X_0$: Pixels

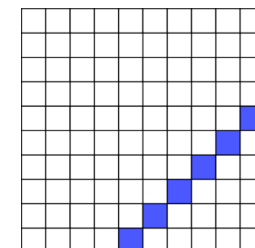
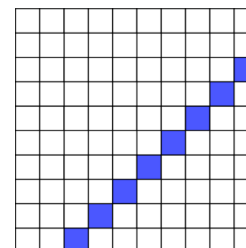
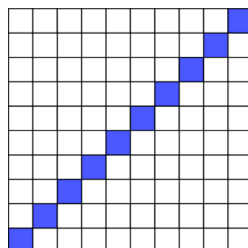
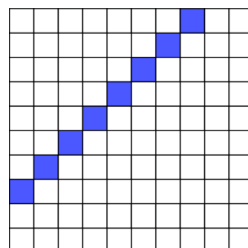
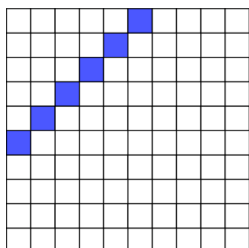
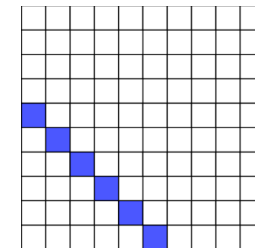
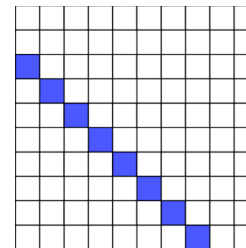
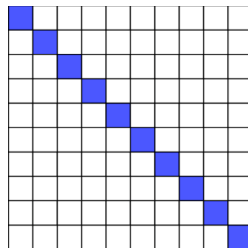
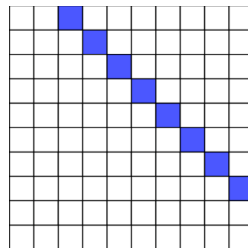
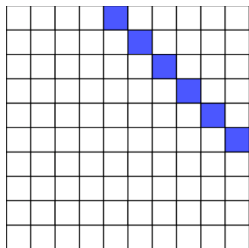
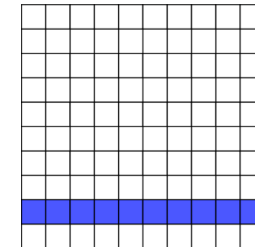
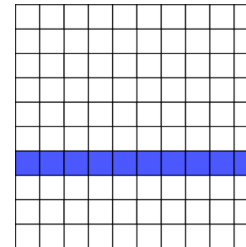
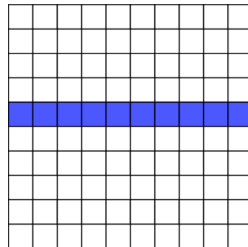
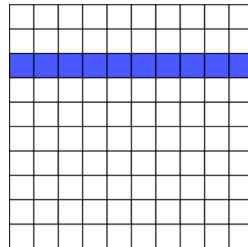
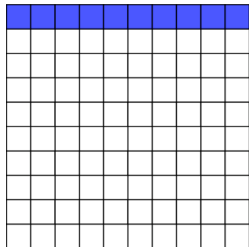
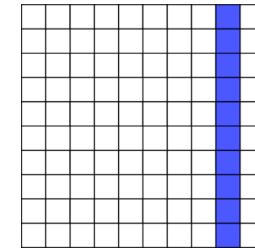
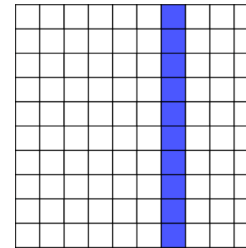
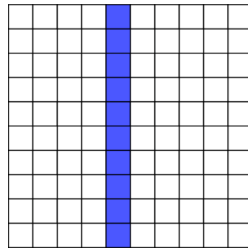
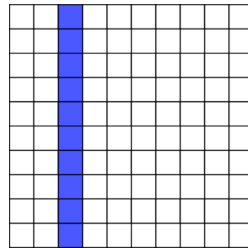
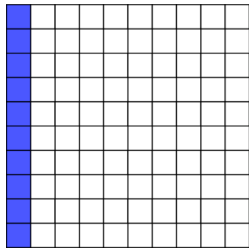
$X =$

-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	Not used
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	Tags
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	Tags
1	1	1	1	1	1	1	1	1	1	Pixels
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	Pixels
.....										

$\};$

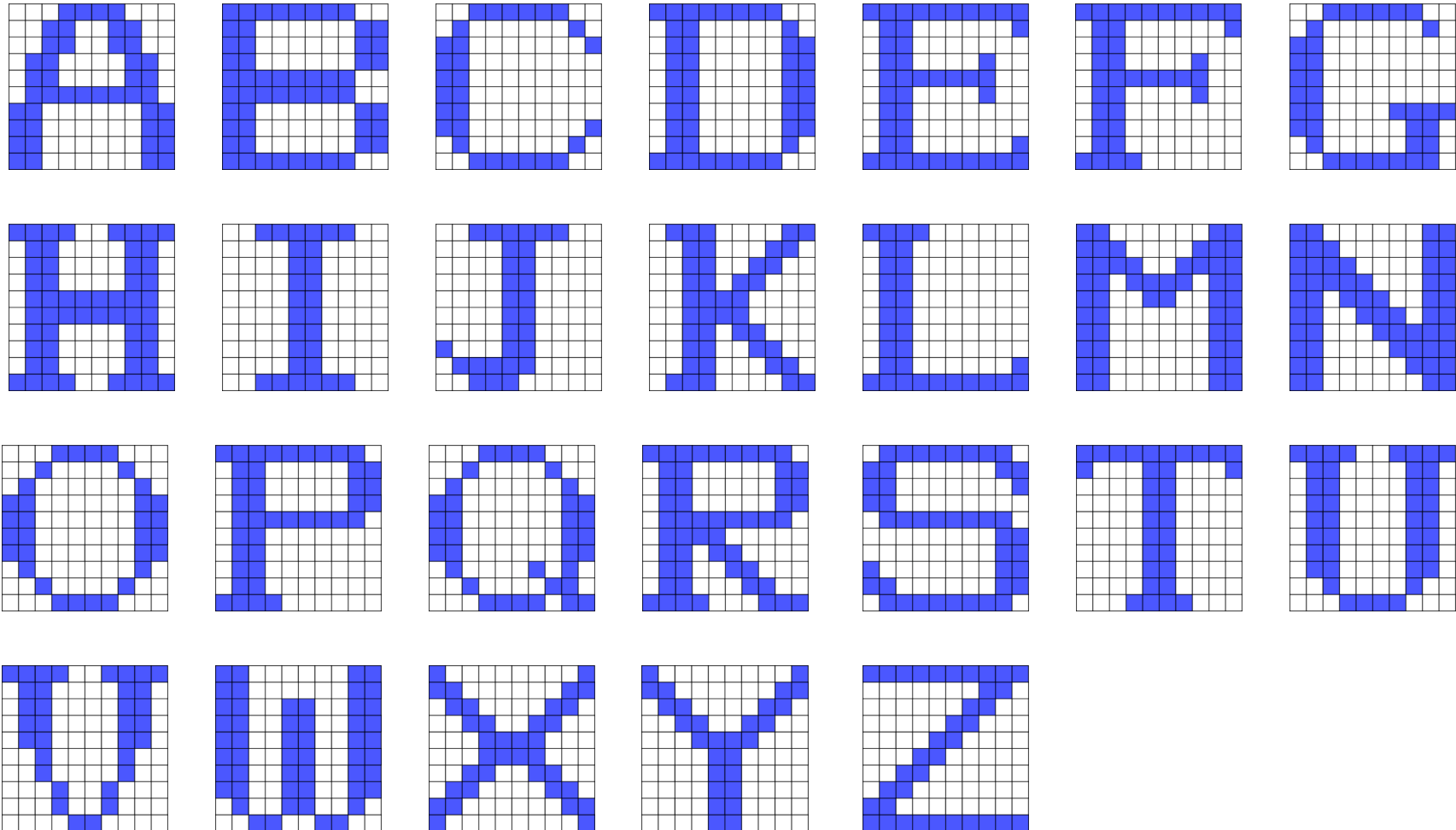


Application 1: Line Patterns



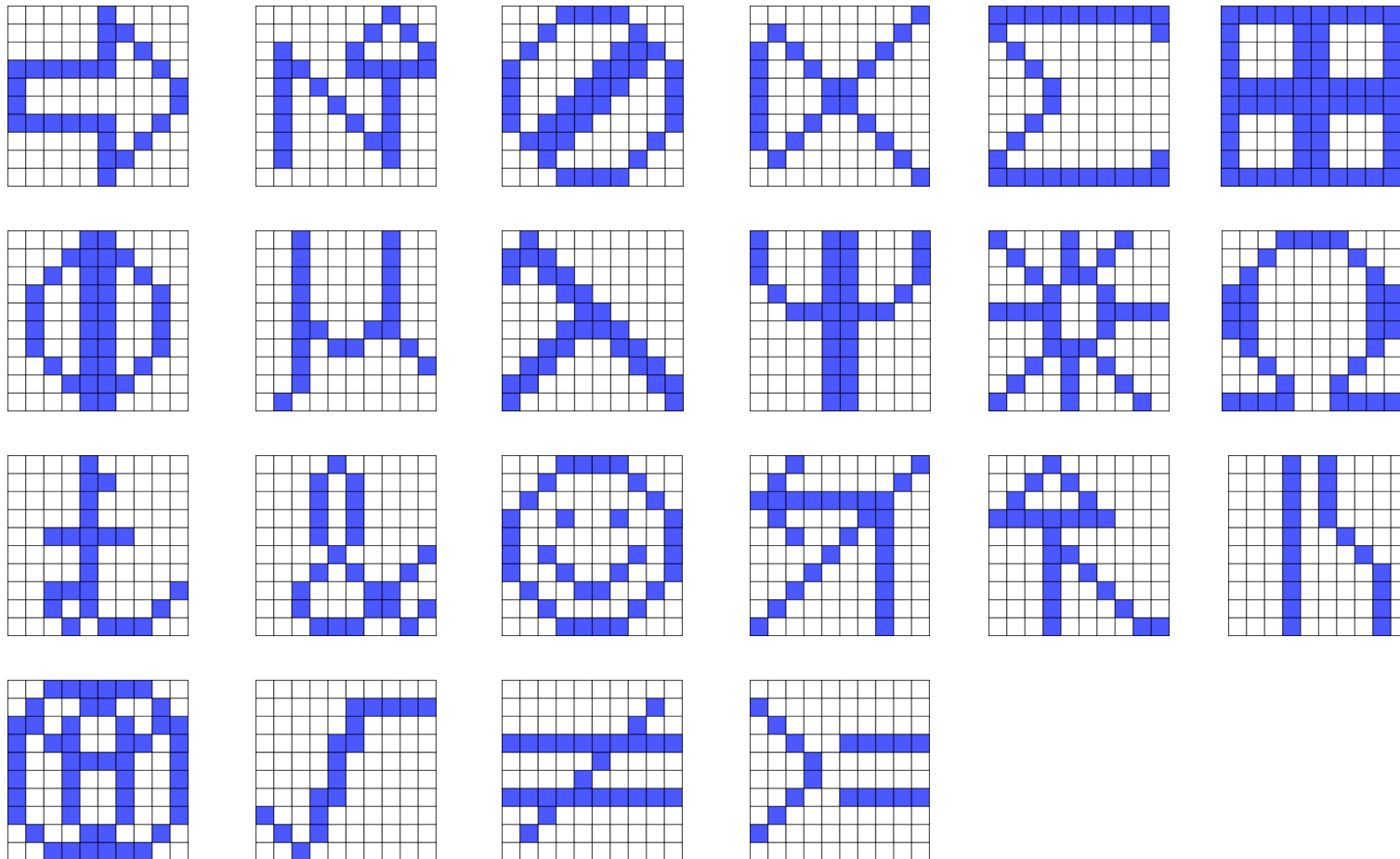


Application 2: Alphabetic Patterns





Application 3: Symbol Patterns





Results



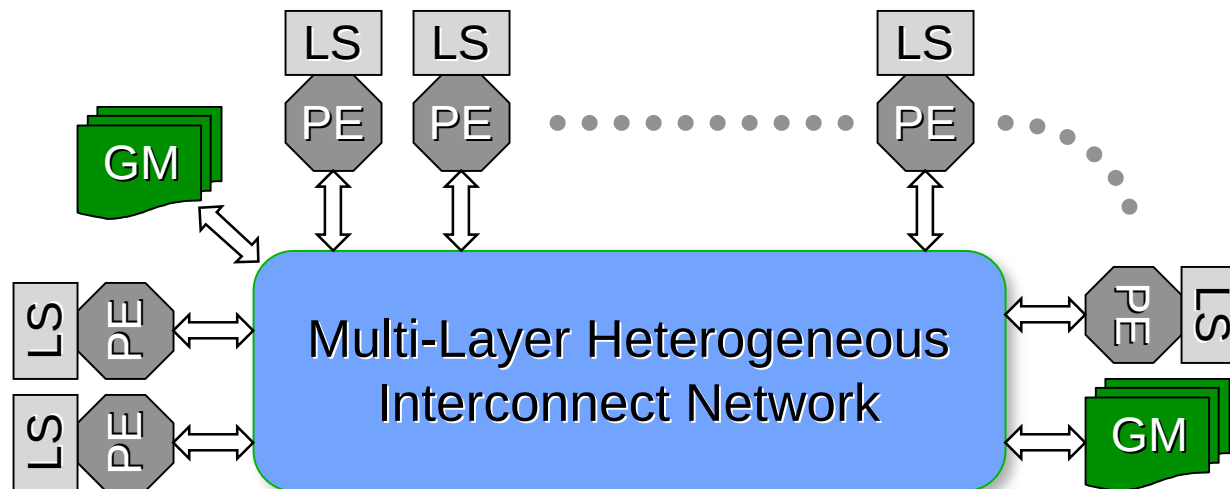
Training: 500 iterations

Recall:

	Number of iterations for successful recalls	Number of iterations for unsuccessful recalls
Exact training pattern without tag	8 ~ 11	N/A
Similar/partial pattern without tag	11 ~ 41	16 ~ 40



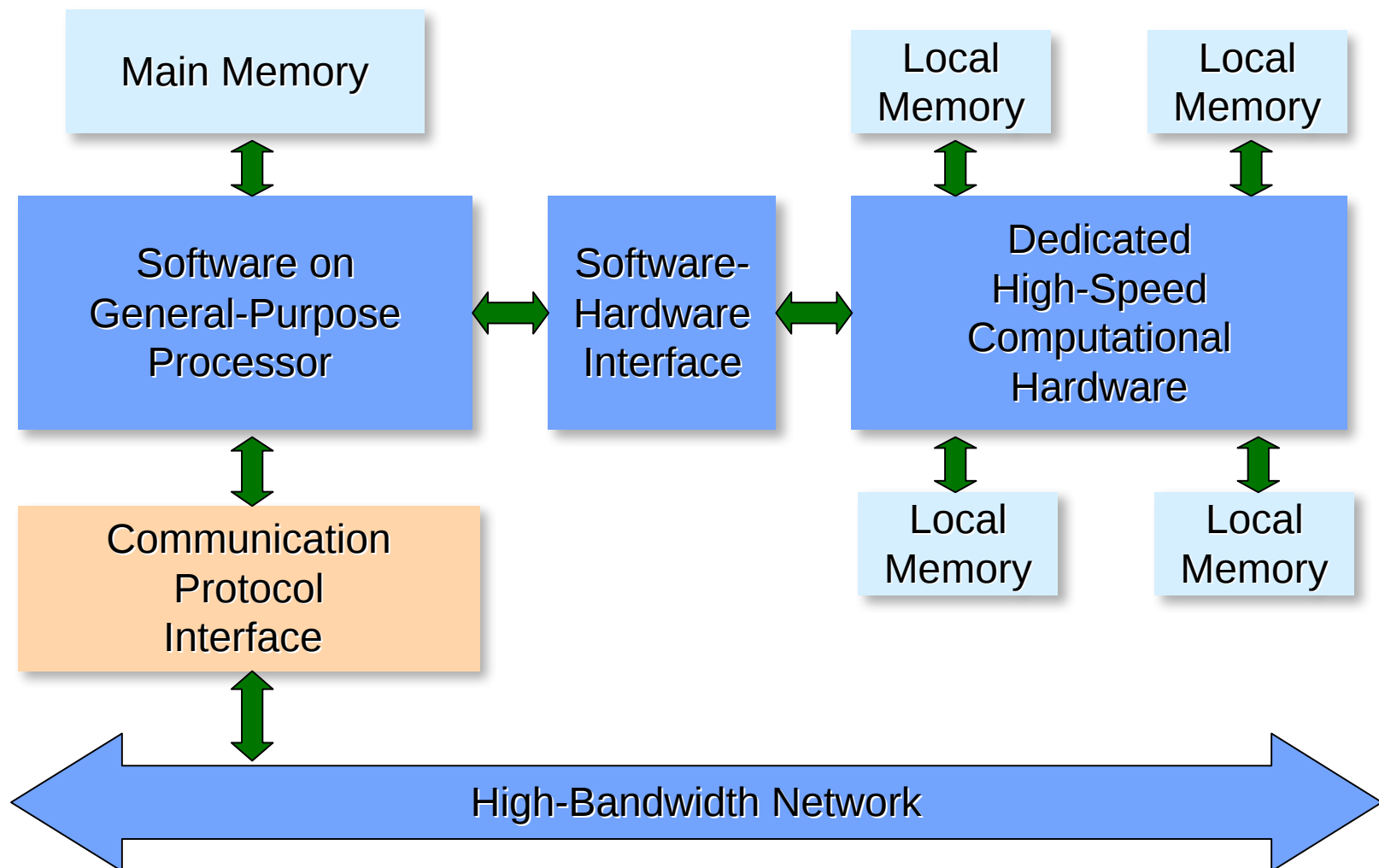
Generic Massively Parallel Cognitive Computing System



PE: Processing Element
LS: Local Store (local memory)
GM: Global Memory

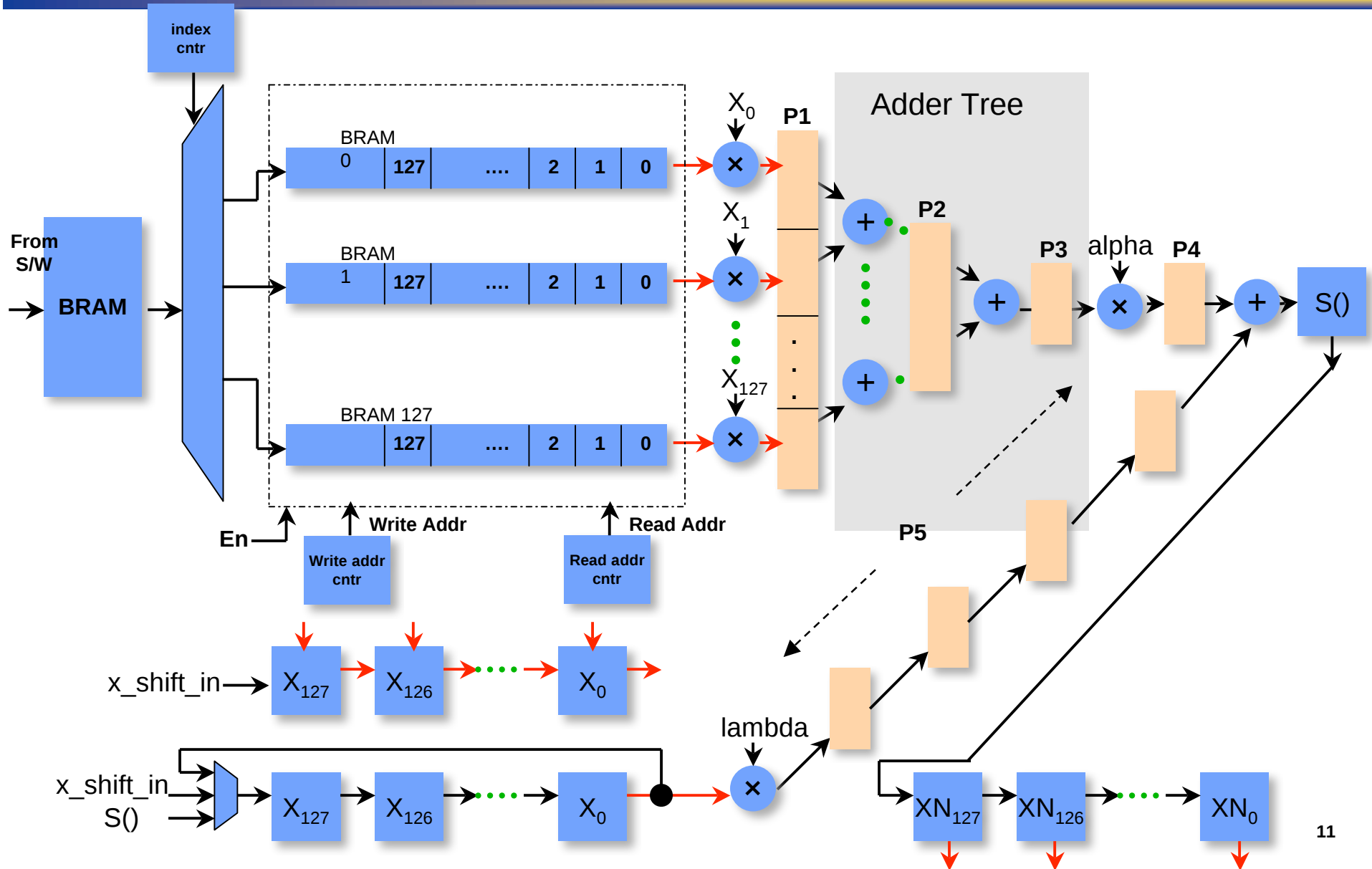


Overall Processed Element Architecture





FPGA Data Path of BSB_128 Recall





FPGA Resource Utilization



FPGA device: Xilinx Virtex II Pro XC2VP70
Target clock frequency: 100 MHz

Resources	Recall Design	Training Design
Block Rams	129 (of 288)	129 (of 288)
18-bit Multipliers	130 (of 328)	130 (of 328)
LUTs	6127 (8%)	22716 (33%)



FPGA Results and Comparison



Software: 2.0GHz Intel Xeon processor, 2GB RAM
Programmed and compiled with Intel Math Kernel Library 9.0

Hardware: 90 MHz clock frequency

Implementation	Time per Recall *(μ s)	Equivalent FLOPS
Software	12.5	~2.6G
Hardware	1.69	~19.4G

* Average of 1,000,000 recalls

Speed up by hardware: 7.4X



Software Solution on PS3



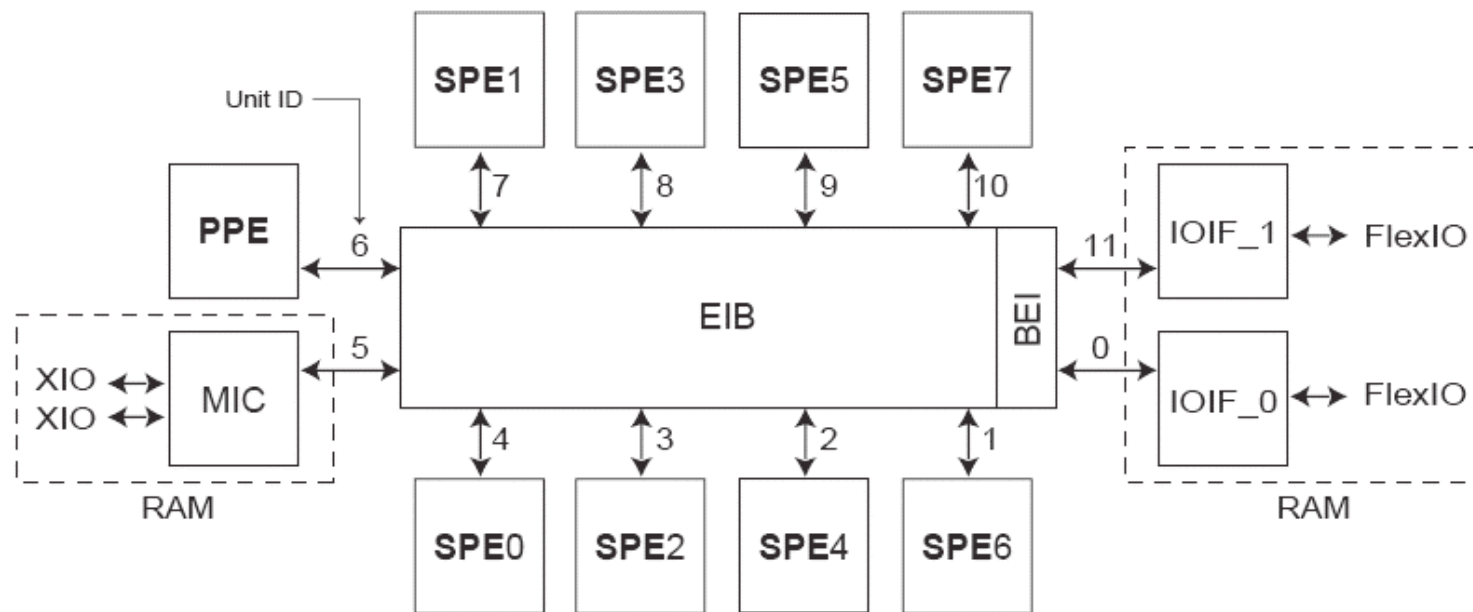
PlayStation 3



- Yellow dog Linux 5.0
- IBM CELL SDK 2.0
- Cell processor
- 256 MB RAM
- 60 GB hard drive
- Gigabit Ethernet
- 150 Gflops Single Precision Peak
- \$499



Cell Broadband Engine Processor

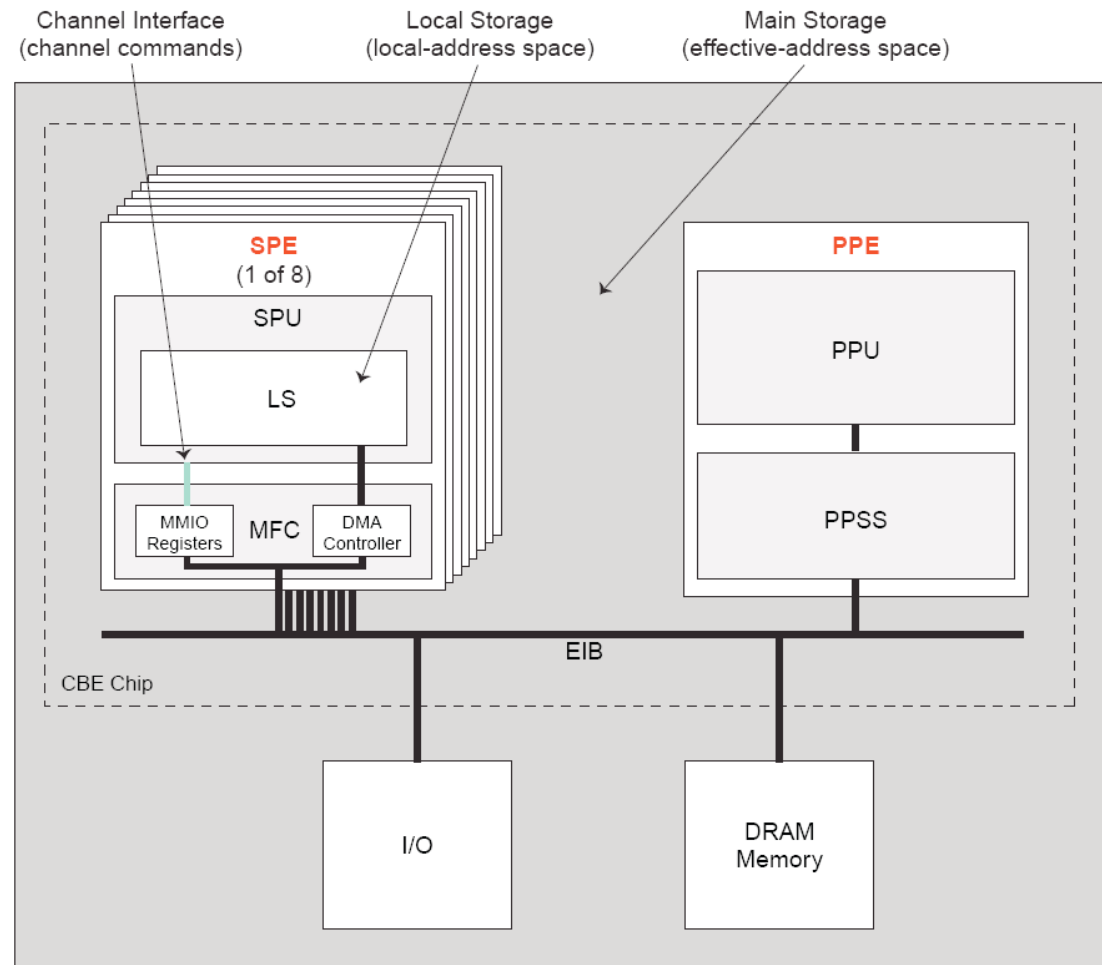


BEI Cell Broadband Engine Interface
EIB Element Interconnect Bus
FlexIO Rambus FlexIO Bus
IOIF I/O Interface

MIC Memory Interface Controller
PPE PowerPC Processor Element
RAM Resource Allocation Management
SPE Synergistic Processor Element
XIO Rambus XDR I/O (XIO) cell



Storage and I/O Interface



DMA Direct Memory Access
EIB Element Interconnect Bus
LS Local Storage
MFC Memory Flow Controller
MMIO Memory-Mapped I/O

PPE PowerPC Processor Element
PPSS PowerPC Processor Storage Subsystem
PPU PowerPC Processor Unit
SPE Synergistic Processor Element
SPU Synergistic Processor Unit



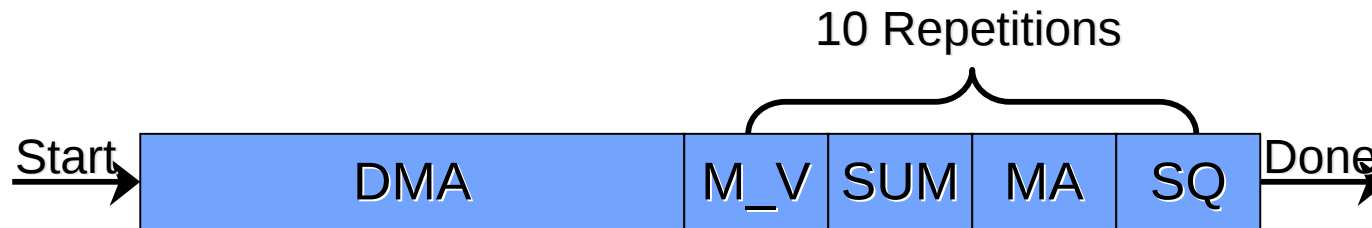
Key Performance Numbers



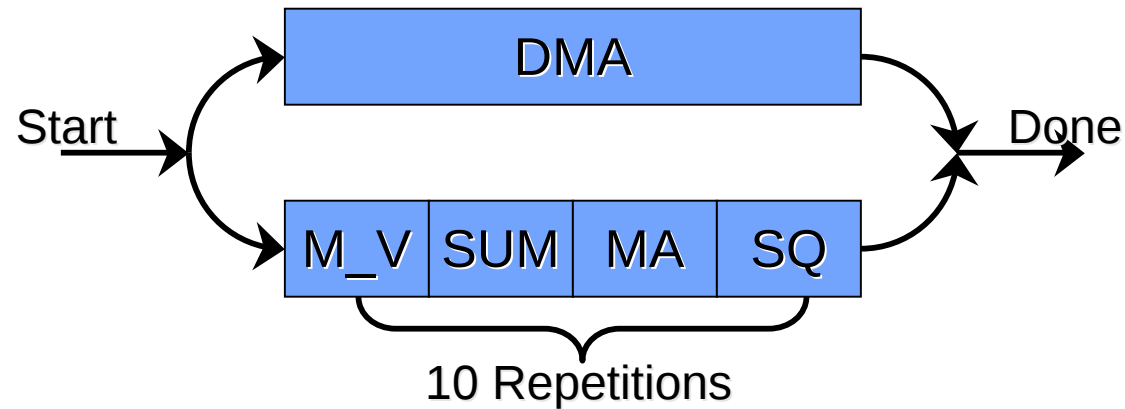
- **Clock Frequency**
 - **3.2 GHz**
- **Peak Performance**
 - **25.6 GFLOPS per SPE**
- **Main Memory**
 - **256 MB Rambus XDR in PS3 and blades**
 - **25.6 MB/s total memory bandwidth**



Computation Flow



(a) Without double buffering



(b) With double buffering

DMA: Fetch coefficient matrix (128x128) from main memory to local store

M_V: Multiplication of coefficient matrix and state vector

SUM: Summation of partial products

MA: Multiply by Alpha and add state vector

SQ: Squash function



Performance Results

100,000 iterations on each SPE

Algorithm Configuratione					Runtime (s)	Performance (GFLOPS / SPE)
M_V	SUM	MA	SQ	DMA		
✓	✓	✓	✓	✓	3.22	10.3
✓	✓	✓	✓		2.32	14.3
✓	✓	✓			2.20	15.0
✓	✓				2.14	15.3
✓					1.79	18.1



Further Optimization

- **Memory I/O**
 - Only reached 12.2 GB/s in previous results
 - **Solution:** Reconfigured and recompiled Linux kernel to support large memory page size
 - **Result:** 24.4 GB/s memory I/O speed; 14.1 GFLOPS per SPE
- **SPE programming**
 - Operations in “SUM” step is not vectorizable.
 - **Solution:** New algorithm that eliminates the “SUM” step.
 - **Goal:** 18 GFLOPS per SPE



Conclusion



- **Brain-state-in-a Box (BSB) cognitive models have been optimized for both FPGA and Cell implementations**
- **>19 GOPS demonstrated on 6M gate Virtex 2 FPGA on Heterogeneous HPC at Rome**
- **> 14 GFLOPS (55% of peak) demonstrated so far on each SPE of Cell BE**
 - **85 GFLOPS on the 6 SPEs of the Cell in a PS3**
 - **170 TFLOPS/\$M price performance at \$499 per PS3**
- **Expect to demonstrate 18 GFLOPS (71% of peak)**
- **Compute/IO ratio of ~20 adequate to balance Cell within PS3 for high performance**
 - **23.4 ops/IO balances 150 GFLOPS to 25.6 GB/sec RDRAM**
- **FPGAs hard pressed to match the price or performance of the Cell**