



# From BEE to BEE2

*Development of Supercomputer-in-a-Box*

Chen Chang, John Wawrzynek, Bob Brodersen  
*University of California, Berkeley*  
*Berkeley Wireless Research Center*



# Outline

---

- Review of BEE system
- BEE2 system
  - Applications
  - Hardware Architecture
  - Design Methodology
- Project Timeline



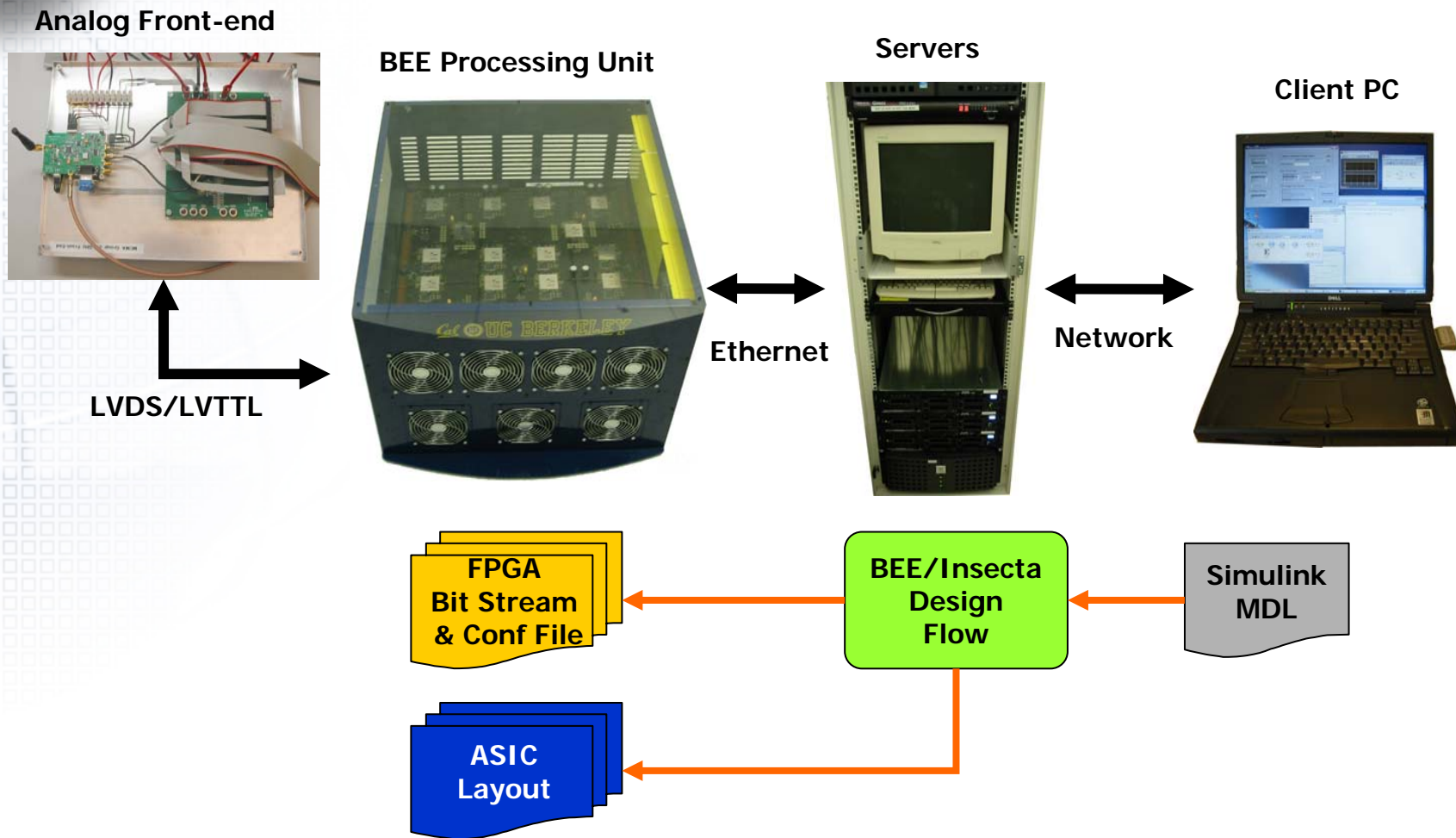
# Berkeley Emulation Engine

- Designed in 2001
- In service at BWRC since March 2002
- FPGA-based system for real-time hardware emulation:
  - Emulation speeds **up to 60 MHz**
  - Emulation capacity of **10 Million** ASIC gate-equivalents, corresponding to **600 GOPS** (16-bit adds)
  - **2400** external parallel I/Os providing **192 Gbps** raw bandwidth.



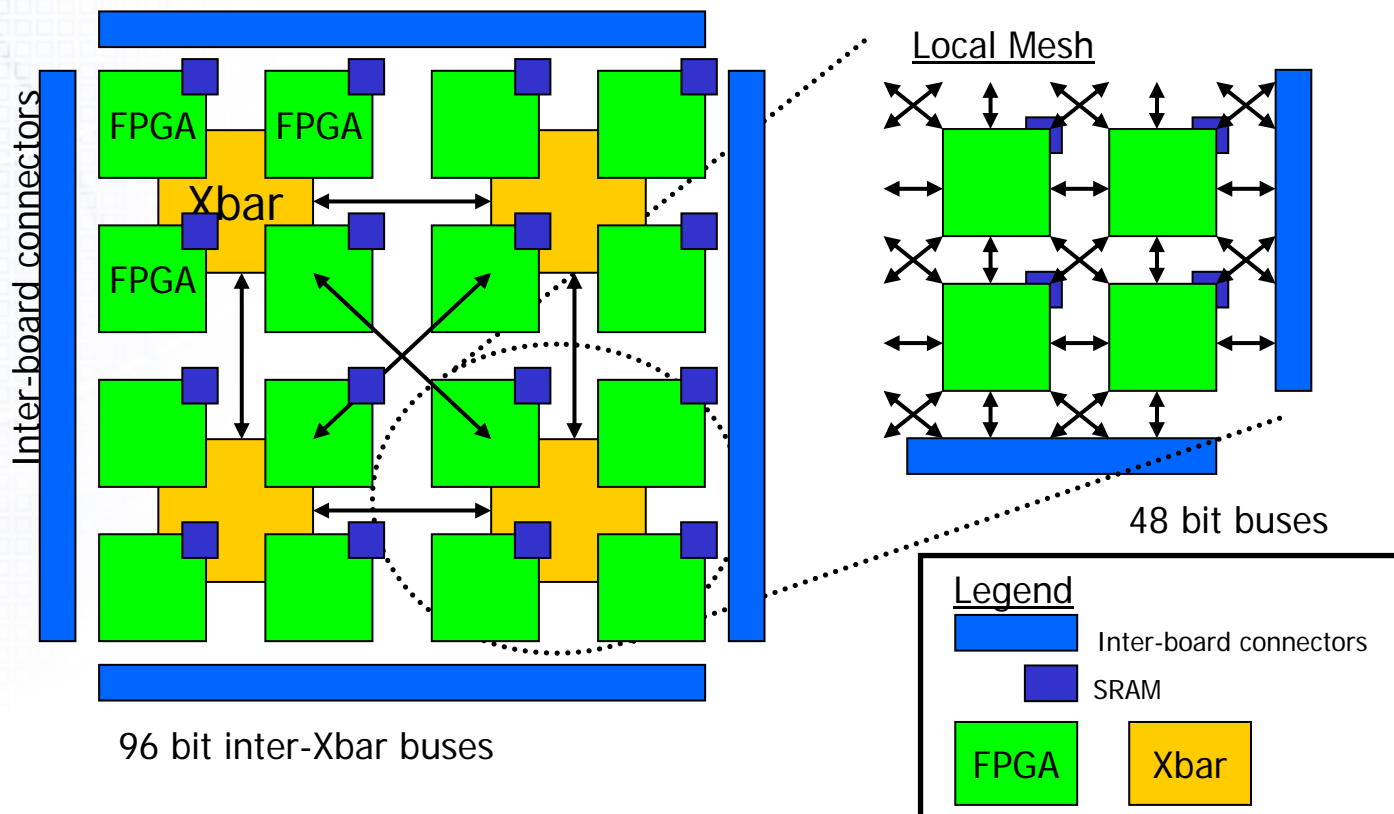


# BEE System Overview





# Virtual FPGA on a PCB







# BEE project results

- Four BEE processing units built
- Developed the Simulink based design methodology targeting both FPGA and ASIC
- Developed a database-driven PCB schematic design framework for large PCB design
- Successful tapeout of:
  - 3.2M transistor pico-radio chip, 1.8M transistor LDPC decoder chip
- System emulated:
  - QPSK radio transceiver, BCJR decoder, MPEG IDCT, etc.
- On-going projects
  - UWB radio, Pico radio multi-node system, Infineon SIMD processor for SDR, Cognitive radio system, etc.



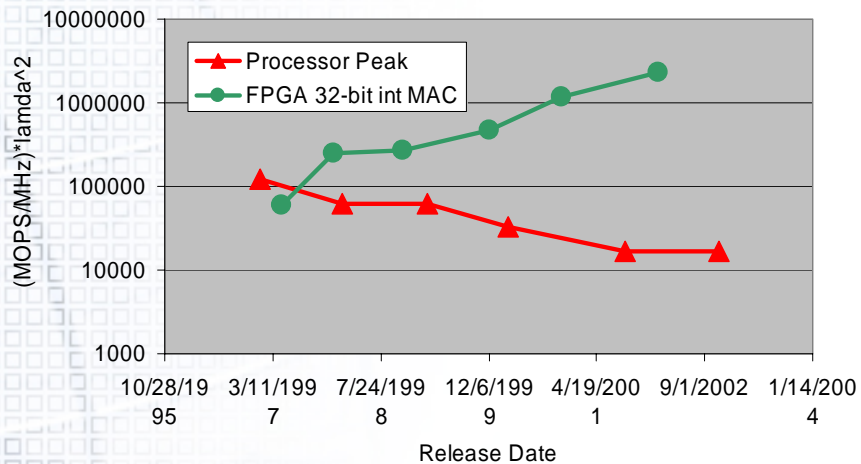
# Lessons from BEE

- Block diagram based dataflow programming model works effectively in the DSP application domain
- Scalability is a key issue
  - BEE PCB board reached the limits of cost, yield, and reliability
  - Off-board parallel I/O interconnect has many issues
    - Signal bandwidth, cable length, noise coupling, mechanical reliability
  - Design methodology directly limits the utilization of the hardware system

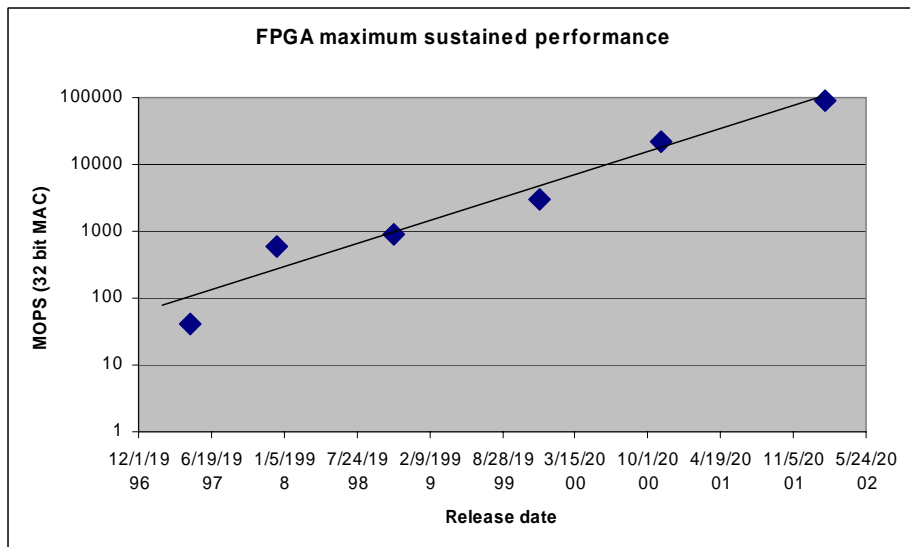


# Moore's Law in FPGA world

Computational Density Comparison



**3X improvement per year!**

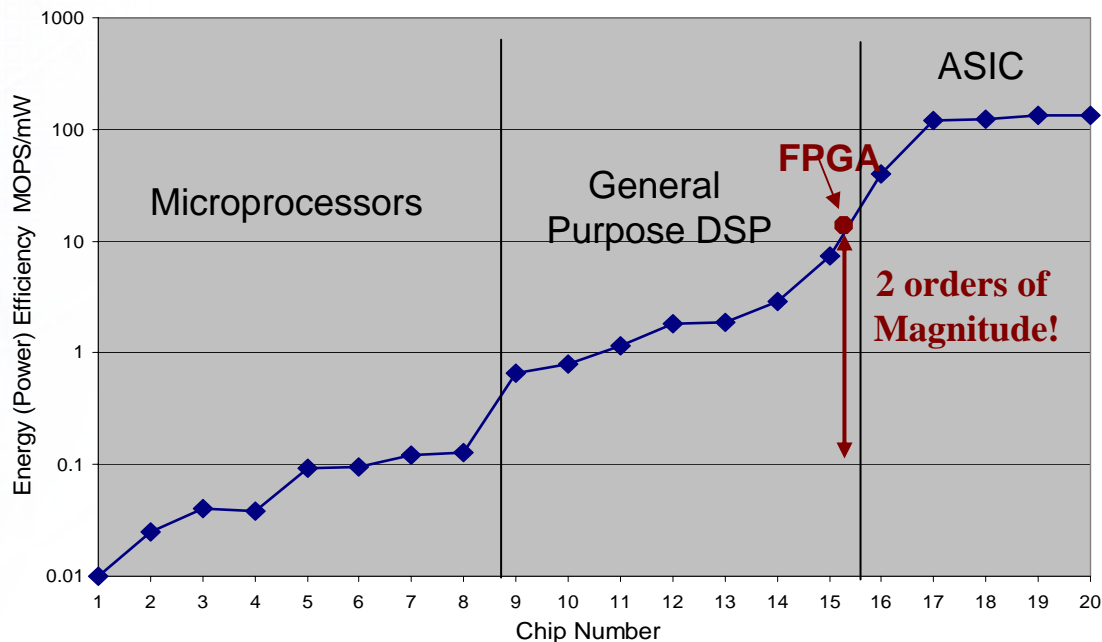






# Energy Efficiency (MOPS/mW or OP/nJ)

- Based published results at ISSCC conferences and our measured results (FPGA).



- Specialized circuits use less energy per operation.
- Inherent computation density means devices can run at lower speed consuming less power.
- Reduced power consumption is a priority for FPGA vendors.



# High-End Reconfigurable Computer (HERC)

- A computer with *supercomputer-like performance*, based solely on FPGAs and/or other reconfigurable devices as the processing elements.
  - The inherent fine-grain flexibility of the FPGAs allow all data-paths, control, memory ports, and communication channels to be customized on a per-application basis and parallelism to be exploited at all levels.
- BEE2 development is underway:
  - demonstrate of the concept;
  - It will motivate engagement of application domain experts;
  - Motivate creative thinking in software/programming tools;
  - Be the first in a series of machines;
  - It is a joint project with Xilinx using their 130nm devices.
- Based on concepts demonstrated in BEE2 prototype, 1 petaOPS ( $10^{15}$ ) in 1 cubic meter attainable within 3 years.



# Applications Areas of Interest

- **High-performance DSP**
  - SETI Spectroscopy, ATA / SKA Image Formation
  - Hyper-spectral Image Processing (DARPA)
- **Scientific computation and simulation**
  - E & M simulation for antenna design (BWRC)
  - Fusion simulation (UW)
  - Musical sound synthesis - finite element models of pianos, etc.
- **Communication systems development Platform**
  - Algorithms for SDR and Cognitive radio
  - Large wireless Ad-Hoc sensor networks
  - In-the-loop emulation of SOCs and Reconfigurable Architectures
- **CAD acceleration**
  - Full Chip Transistor-Level Circuit Simulation (Xilinx)
  - FPGA Place & Route
- **Bioinformatics**
  - BLAST (Basic Local Alignment Search Tool) biosequence alignment .



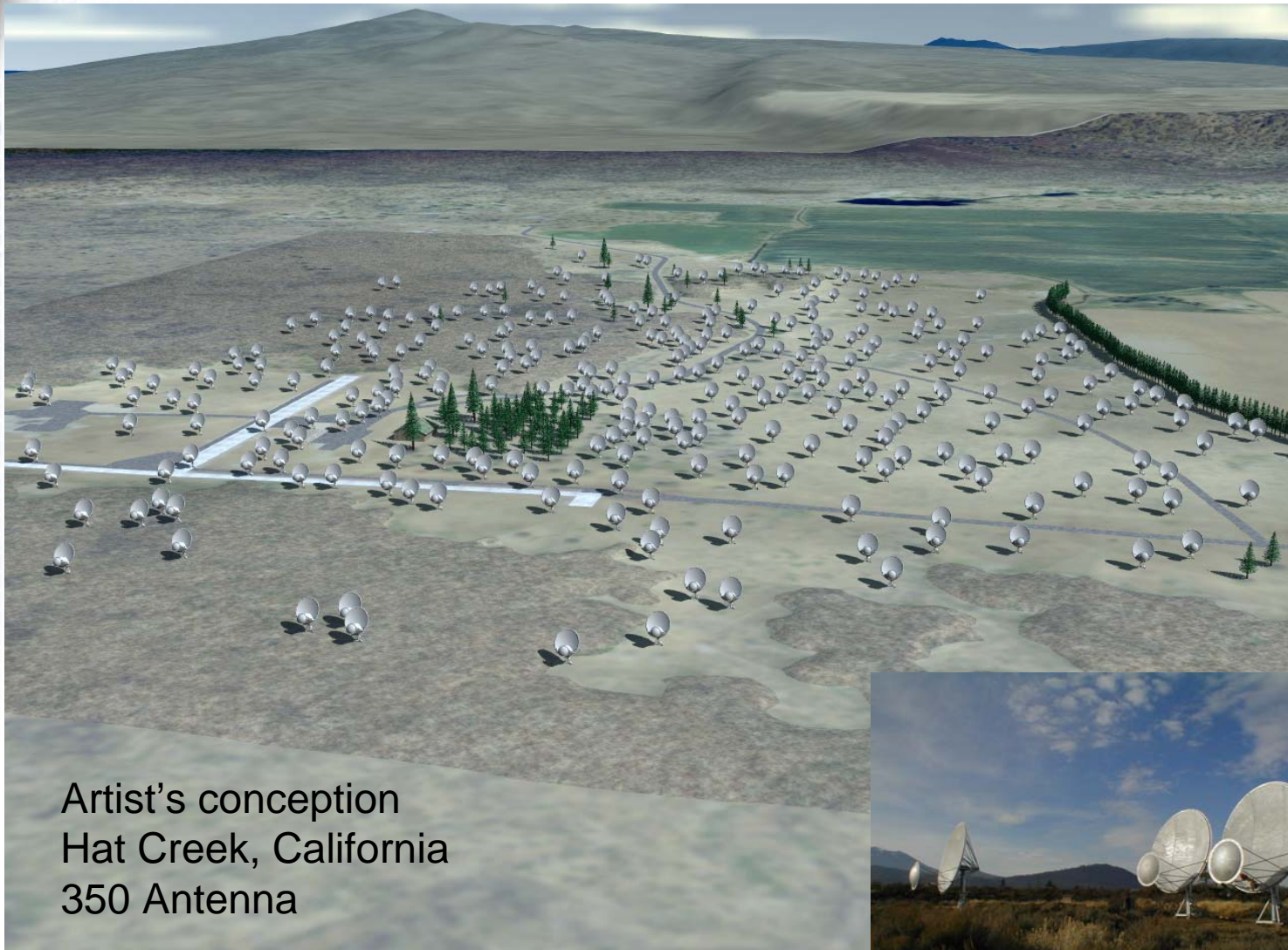
# High-performance DSP

- Large antenna array signal processing
  - Allen Telescope Array (ATA) in 2005
    - Privately funded by Paul Allen, owned by SETI institute
    - 350, 6m antennas, 11GHz RF bandwidth, 0.4GHz IF
    - Full image formation require  $10^{14}$  OPS
  - Square Kilometer Array (SKA) in 2015
    - Internationally funded, \$1B budget
    - 4400, 12m antennas, 22 GHz RF bandwidth, 4GHz IF
    - Full image formation require  $10^{17}$  OPS





# Allen Telescope Array (CA)



Artist's conception  
Hat Creek, California  
350 Antenna



Dec 2nd, 2004

BWRC, UC Berkeley



# Wireless-Network Simulation

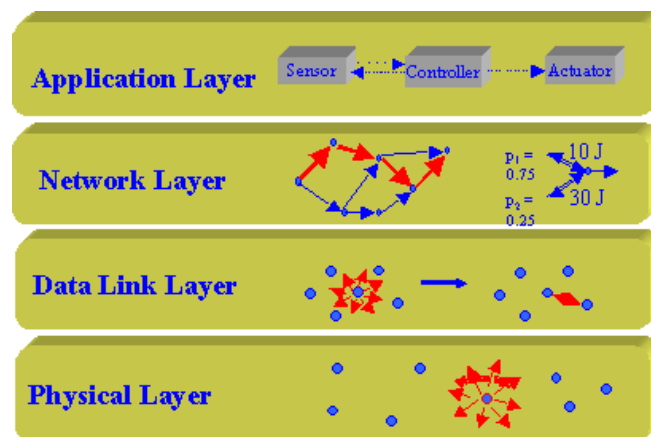
- SDR, Cognitive radio, and Ad-hoc Networks

- Platform for developing soft-radio techniques,
- validation of network protocols,
- chip-level validation in context of real data and network/environment.

- Requires real-time:

- Simulation of complex channel and environment models,
- Simultaneous simulation of 100's to 1000's of network nodes (with real-time sensor input).

*Model for future single-chip reconfigurable computing architectures.*

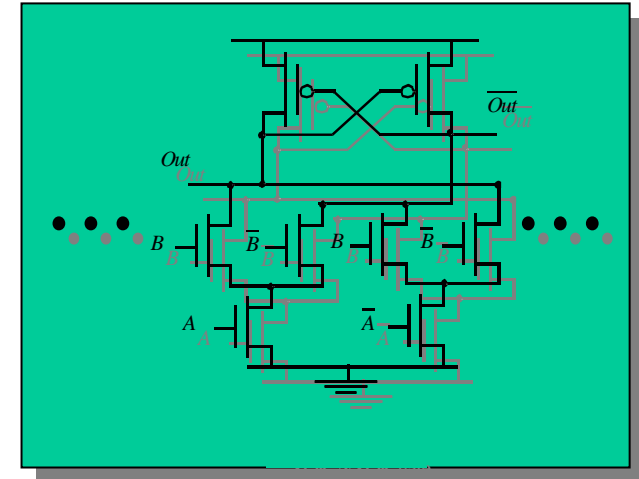






# Rethinking SPICE-level Circuit Simulation

- Conventional implementation:
  - turns circuit into a large 2D conductance matrix (representing connections between nodes with circuit elements).
  - “Gauss elimination” like solver is used to “solve” the matrix for voltages.
  - Requires global communication and floating-point arithmetic.

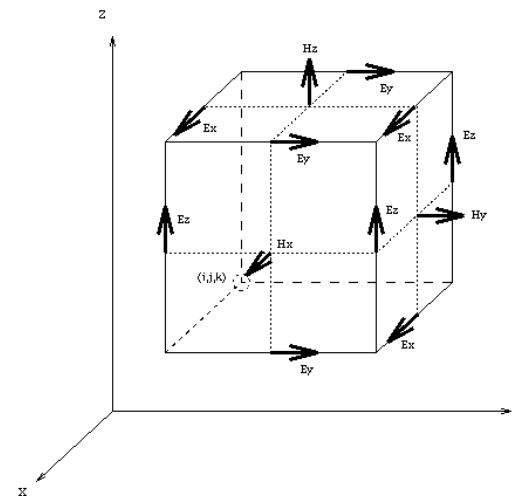


- On BEE2:
  - Circuit is mapped spatially across the computation elements, each responsible for updating a set of nodes iteratively.
  - Communication is localized and fixed point arithmetic suffices.



# FDTD E&M Simulation

- Typical problem has  $500^3$  grid points and 10,000 time step simulation:
  - ~ 20 hours on a workstation
- Yee cell engine:
  - 21K LUTs, 6.4GB/s @ 120MHz
  - Uses FP units from eda.org.
- Single FPGA (V2P70/100):
  - With 3 engines per chip,
  - 1.9 hours
- On BEE2, less than a minute.

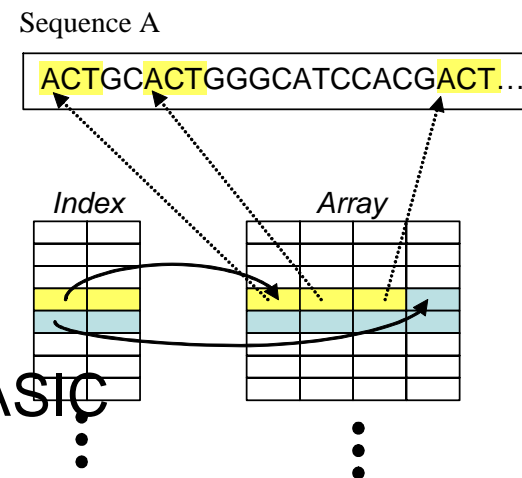


$$E_x = k_{E1} \times E_x + k_{E2} \times (H_y - H_y^* - H_z + H_z^*)$$
$$H_x = k_{H1} \times H_x + k_{H2} \times (E_y - E_y^* - E_z + E_z^*)$$



# Bioinformatics

- Implicitly parallel algorithms
  - Often stream-like data processing
  - Integer operations sufficient
- History of success with reconfigurable/ASIC architectures. (TimeLogic, Paracell)
- High-quality “Brute force” Smith-Waterman technique practical on BEE but not on PC clusters.



*BLAST (Basic Local Alignment Search Tool): Preliminary implementation (simulation only) indicates BEE2, provides 100~1000 times faster execution time running the BLAST algorithm, and over 1000X lower price-performance than existing PC cluster solutions.*

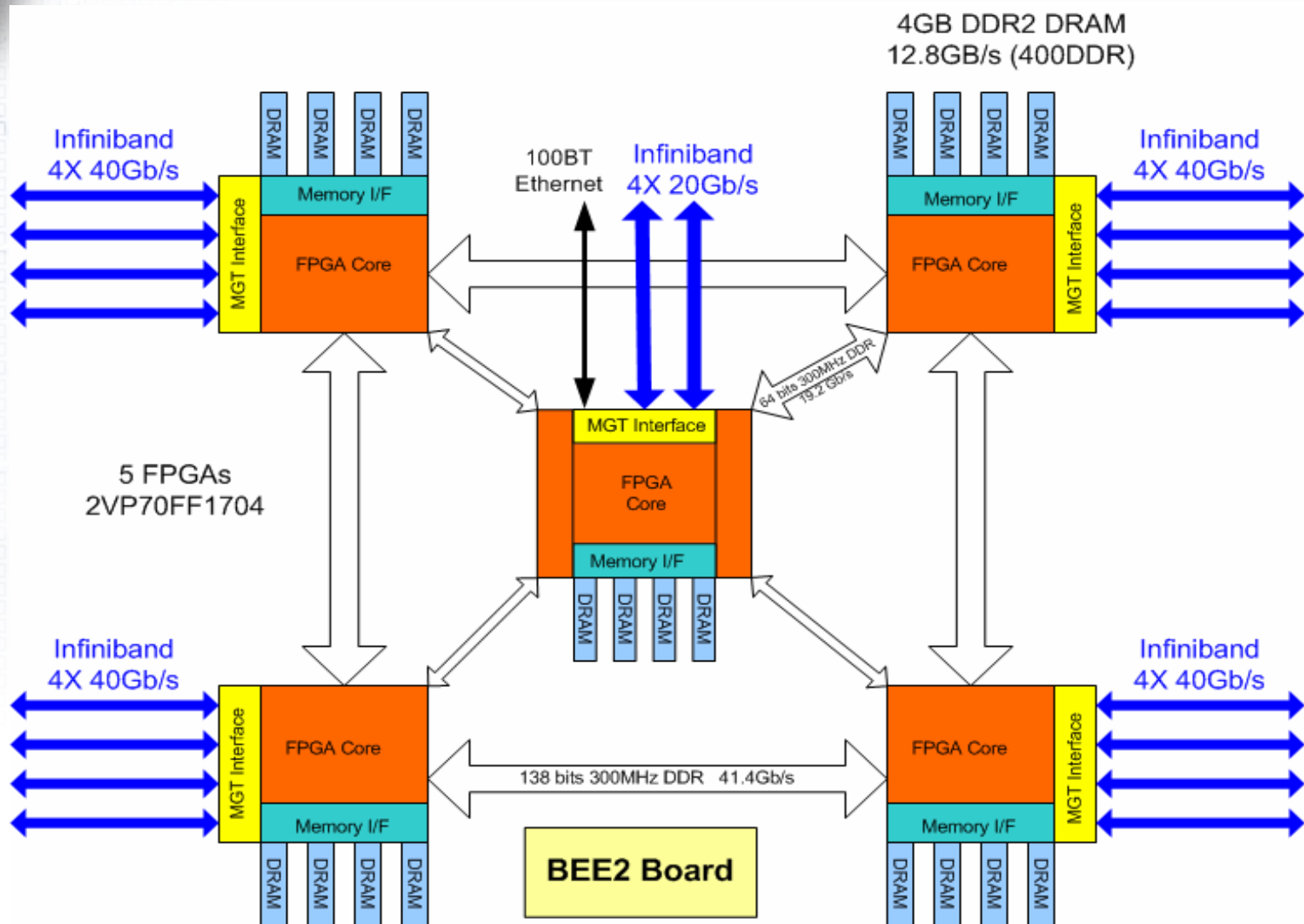


# BEE2 System Design Philosophy

- “Compute-by-the-yard”
  - Modular computing resource
    - Number and interconnection of modules assembled on a per installation basis.
  - Flexible interconnect architecture
  - Per-application reconfiguration of computing resources
- Economy-of-scale
  - Ride the semiconductor industry Moore’s Law curve
  - All COTS components, no specialized hardware
  - Survival of application software using technology independent design flow

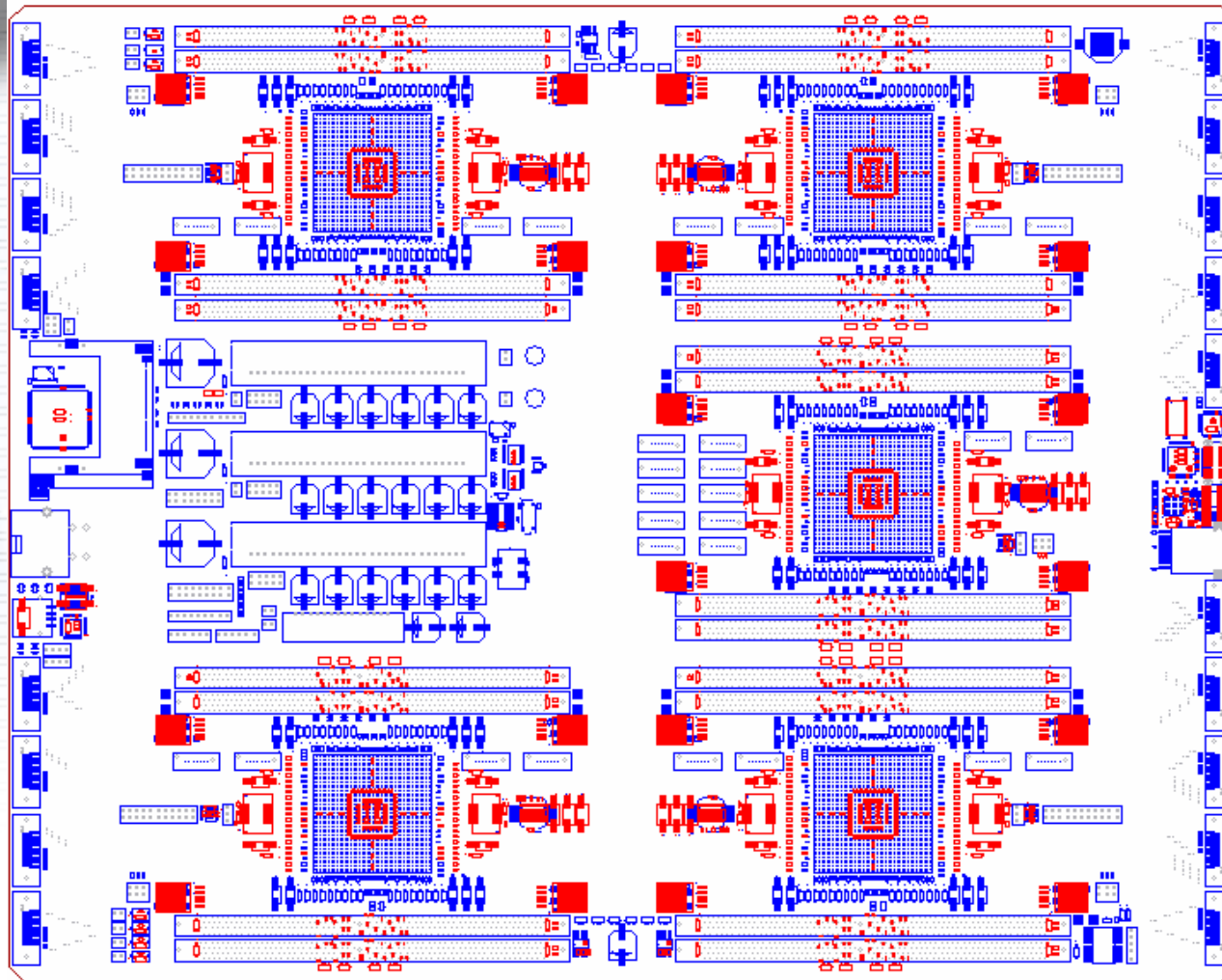


# Compute Module Details





# BEE2 Module: board layout



Module also includes I/O for administration and maintenance:

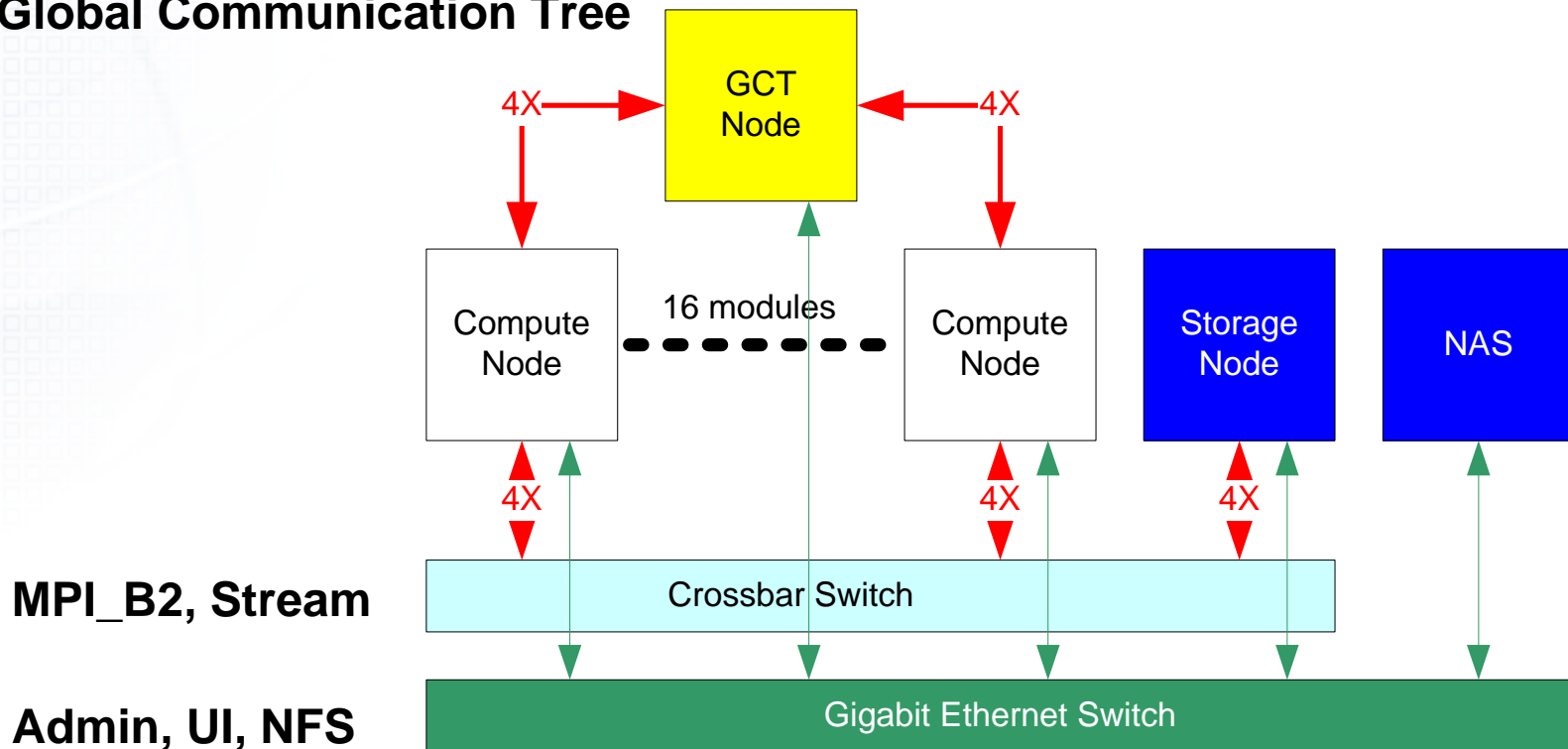
- 10/100 Ethernet
- HDMI / DVI
- USB





# Internode Connections

## Global Communication Tree



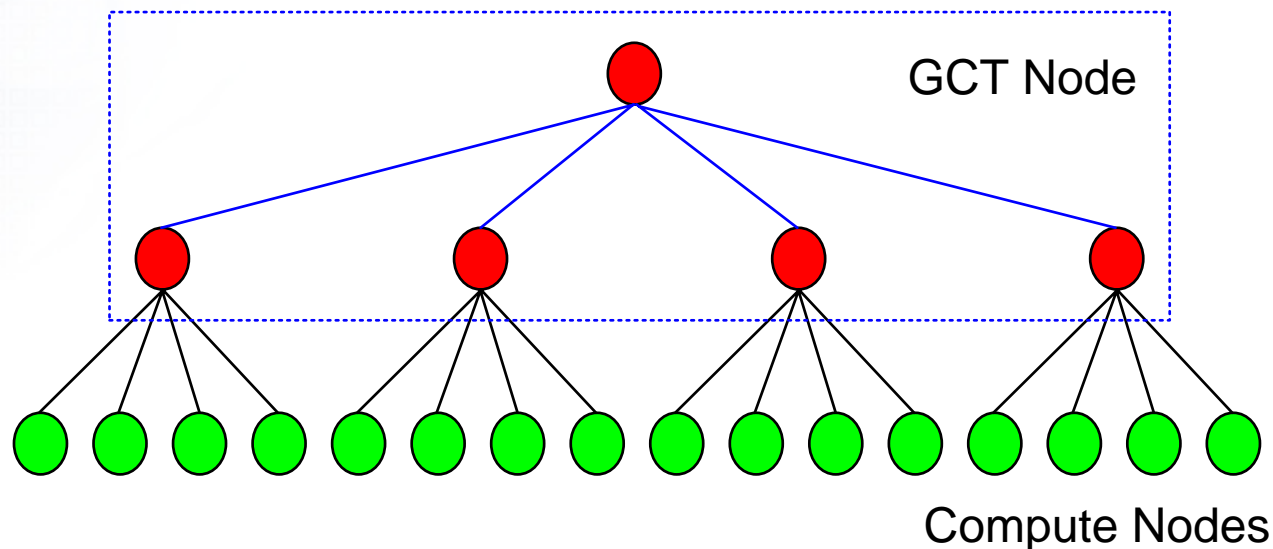
**MPI\_B2, Stream**

**Admin, UI, NFS**



# Global Communication Tree

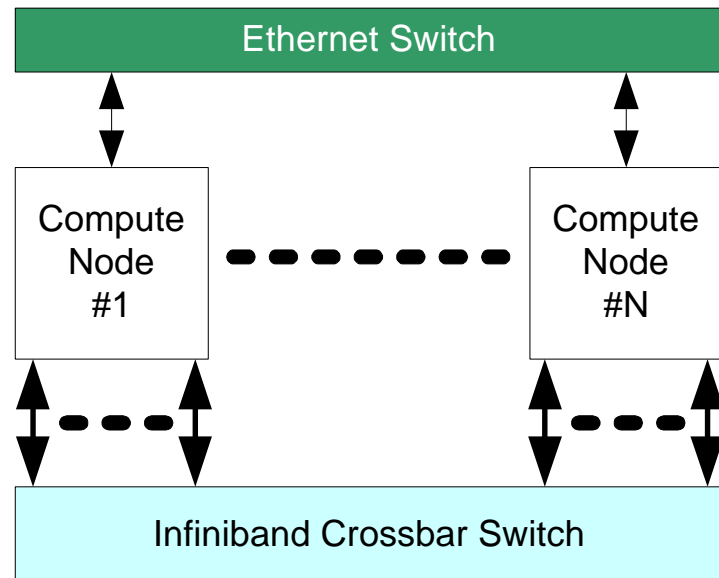
- 4-ary tree configuration
  - Off-module 4X Infiniband connection @ 10 Gbps duplex
  - On-module 64-bit 300DDR connection @ 19.2 Gbps
- Every 16<sup>th</sup> B2 modules act as a tree node
- Each tree node/leaf node has up to 4GB DRAM





# Commercial switches

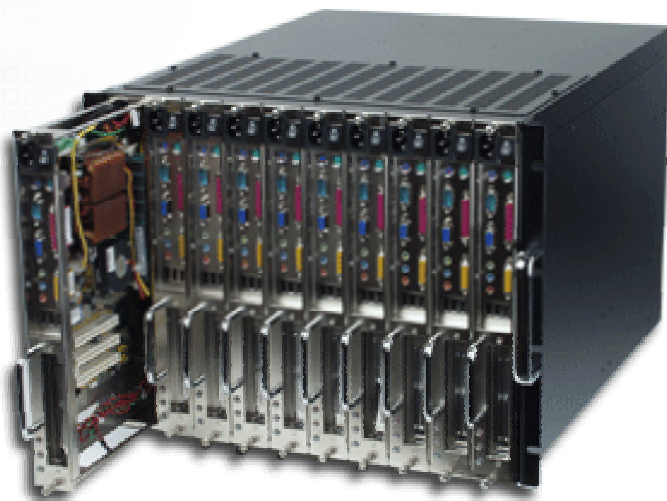
- Commercial Infiniband switch from Mellanox, Voltaire, InfiniCon.
  - Packet switched, non-blocking
  - 24 ~ 288 ports (4X) per chassis
  - Up to 10,000 ports in a system
  - 200~1000 ns switch latency
  - 480Gbps ~ 5.76Tbps full duplex constant cross section bandwidth
  - <\$400 per port
- Ethernet
  - System control, monitoring, debugging
  - Low bandwidth connection to data archiving, and backend processing





# 19" Rack Cabin Capacity

- 40 compute nodes in 5 chassis (8U) per rack
- 32-40TeraOPS, 1.3-1.6TeraFLOPS
- 250 Watt AC/DC power supply to each blade
- 12.5 Kwatt max power consumption
- Hardware cost: ~ \$800K





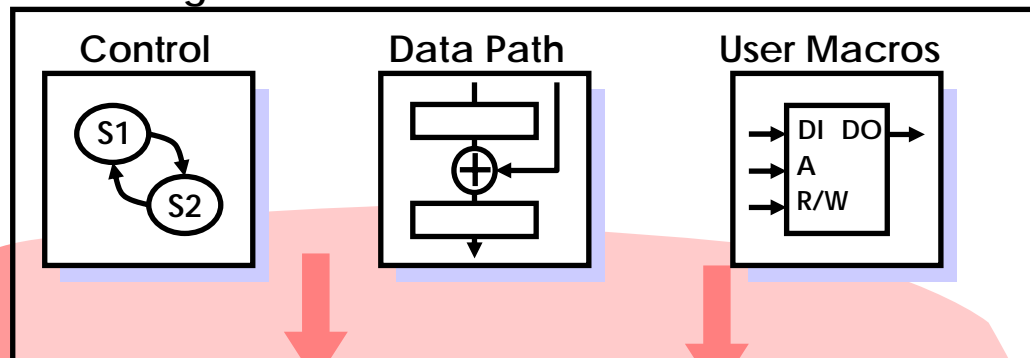
# Existing FPGA Programming Models

- Traditional ASIC approach
  - VHDL/Verilog RTL description of the circuit
  - High performance, versatile.
  - Advanced synthesis knowledge required.
- Software programming approach
  - Compile C/Java/Matlab to RTL HDL then synthesize
  - The original sequential program order limits parallelism hence performance
- Data Flow Diagram + Finite State Machine
  - Targeting digital signal processing applications
  - Minimal performance impact with predefined IP cores



# BEE Programming Model : Discrete Time Block Diagram with FSM

Block Diagrams:



Matlab/Simulink:  
Functional simulation,  
Hardware Emulation

StateFlow,  
Matlab  
HDL

CoreGen  
Module  
Compiler

Black Boxes

- Implementation details of memory and communication interfaces are left in the user domain
- Abstraction level of these interfaces are too low to support efficient design of large-scale systems





# BEE2 hardware abstractions

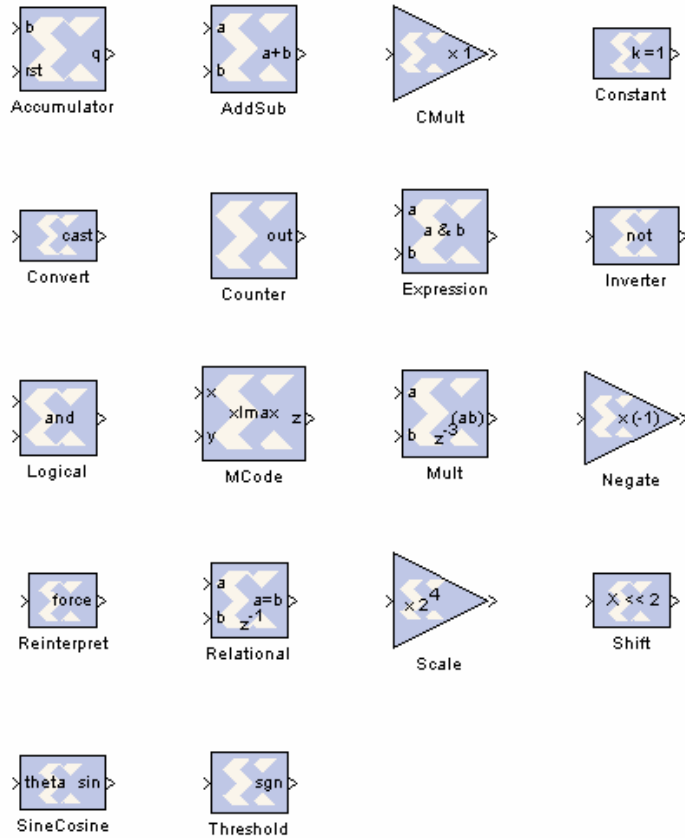
- Data flow operators
  - Data type: fix-point
  - Math operators: +/-, \*, /, &, |, xor, ~, >, =, <, srl, sll, sra
  - Control operators: demux/switch, mux/merge
- Memory
  - On-chip SRAM/Registers: shift register, RAM, ROM
  - Off-chip DRAM: stream RAM
- Communication and I/O
  - Static links: stream I/O
  - Dynamic links: Remote DMA
- Synchronization
  - Time stamp



# Data flow operators

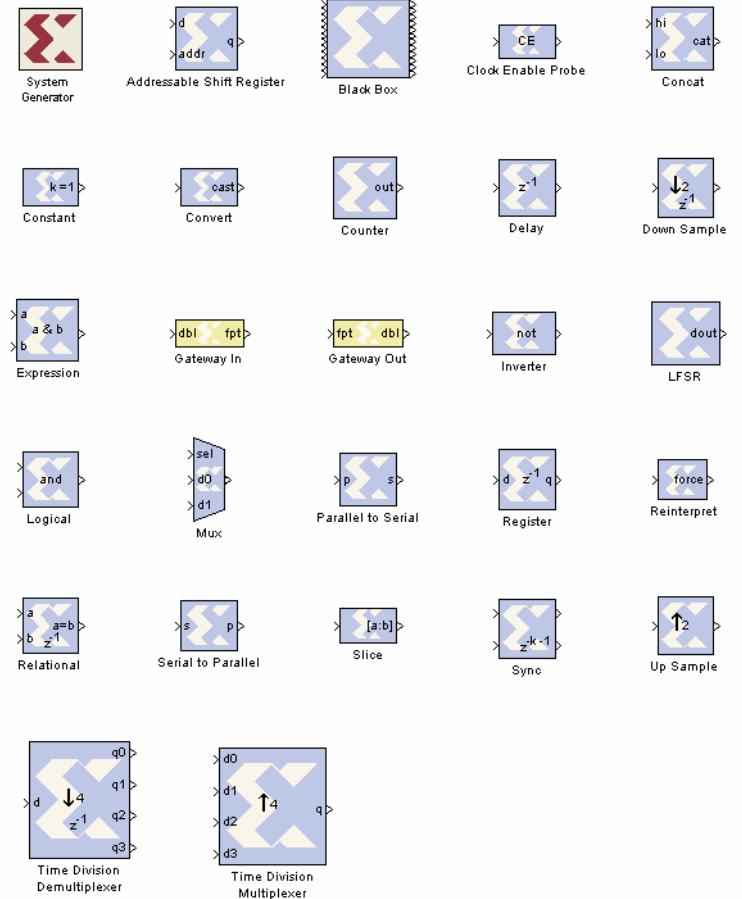
Xilinx Blockset v6.2  
(c) 2004 Xilinx, Inc.

## Math Library



Xilinx Blockset v6.2  
(c) 2004 Xilinx, Inc.

## Basic Library

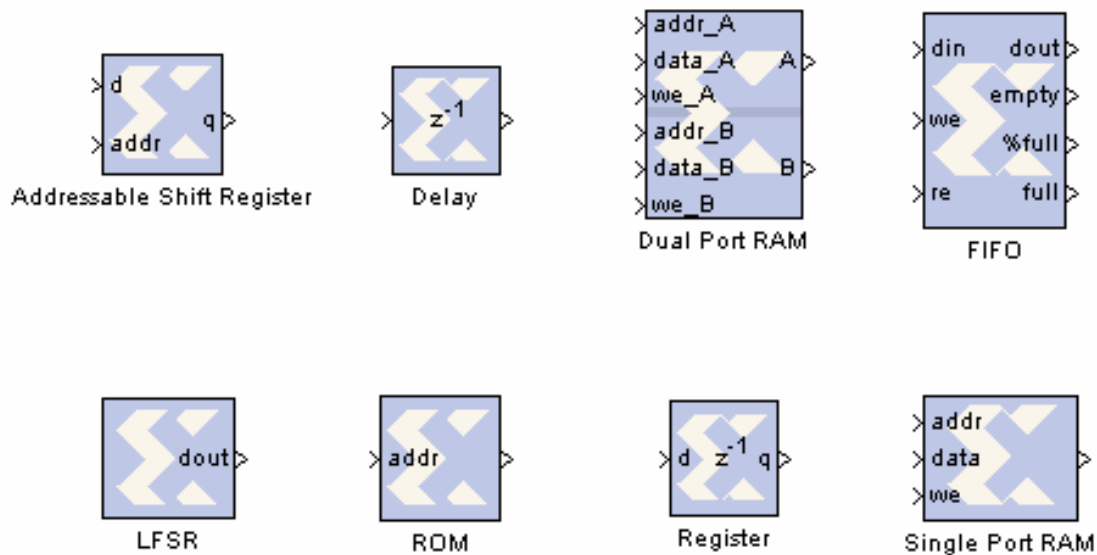




# On-chip memory

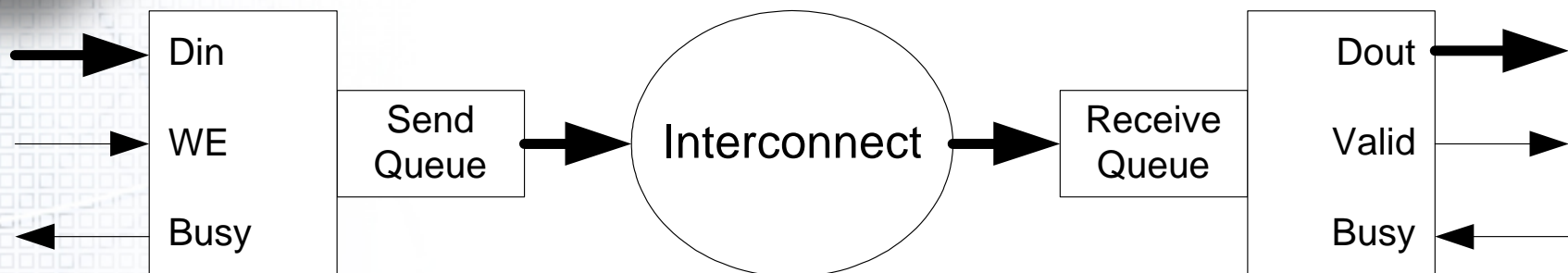
Xilinx Blockset v6.2  
(c) 2004 Xilinx, Inc.

## Memory Library





# Stream I/O abstraction for inter-chip static connections



- Statically allocated system resource
- Guaranteed throughput, but not fixed latency
- Finite queue length, user configurable
- Independent of physical implementation
  - Circuit/packet switch, dedicated link
  - Parallel I/O, MGT, Infiniband



# Staged Development of Peta-BEE

machine	BEE2 prototype	BEE2 full-rack	BEE3 full-rack	BEE4 prototype	BEE4 full-rack
year	Q1 2005	Q3 2005	Q1 2006	Q1 2007	Q3 2007
chip technology	130 nm	130 nm	90 nm	65 nm	65nm
fixed-point perf.	1.6-2 TOPS	32-40 TOPS	128-160 TOPS		<b>2048-2560 TOPS</b>
FP performance	64-80 GFLOPS	1.3-1.6 TFLOPS	5.12-6.4 TFLOPS		82-122 TFLOPS
with acceleration					(400-600 TFLOPS)
special				custom masks	custom masks
				stacked-die	stacked-die

- **Notes:**

- “prototypes” are 2 modules.
- “full-rack” versions are 40 modules, plus necessary switches, power supplies, etc.
- BEE3 implementation is optional.
- Schedule reflects technical feasibility. Development schedule would build in slack.

- **BEE4:**

- Assumes 65nm by late 2006.
- Special masks may be needed to provide proper balance of I/O, memory, and logic.
- Special masks could boost floating point performance (5x) if needed by applications.
- Memory die stacked on FPGAs to gain 4x in density.





# The BEE2 Team

- Faculty in charge
  - *John Wawrzynek*
  - *Bob W. Brodersen*
- Graduate students
  - *Chen Chang*
  - *Pierre-Yves Droz*
  - *Nan Zhou*
  - *Yury Markovskiy*
  - *Zohair Hyder*
  - *Adam Megacz*
  - *Alexander Krasnov*
  - *Hayden So*
  - *Kevin Camera*
- Industrial Liaison
  - *Bob Conn (Xilinx)*
  - *Ivo Bolsens (Xilinx)*
- Research associates
  - *Dan Werthimer (SSL)*
  - *Melvyn Wright (UCB, RAL)*
  - *Don Backer (UCB, astro)*
- Technical staff
  - *Brian Richards*
  - *Susan H. Mellers*
- Undergraduate student
  - *John Conner*
  - *Greg Gibeling*