FPGA, a history of interconnect

Ivo Bolsens

CTO, Xilinx

XILINX

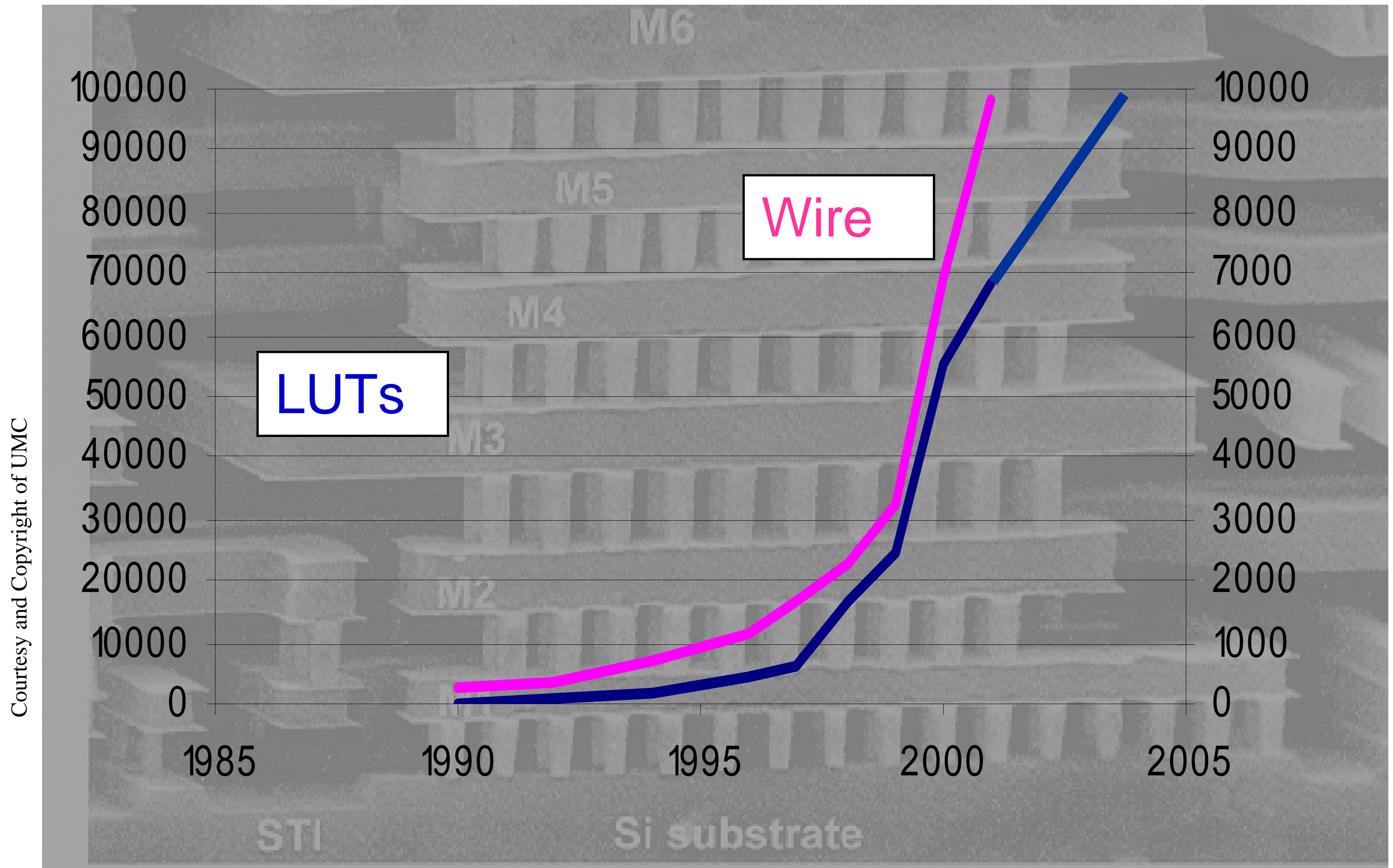
The Architectural Shakeout

Mass extinctions in the mid 1990s

- Xilinx: 8100, 6200, 4700, Prizm, …
- Plessey, Toshiba, Motorola, IBM, ...

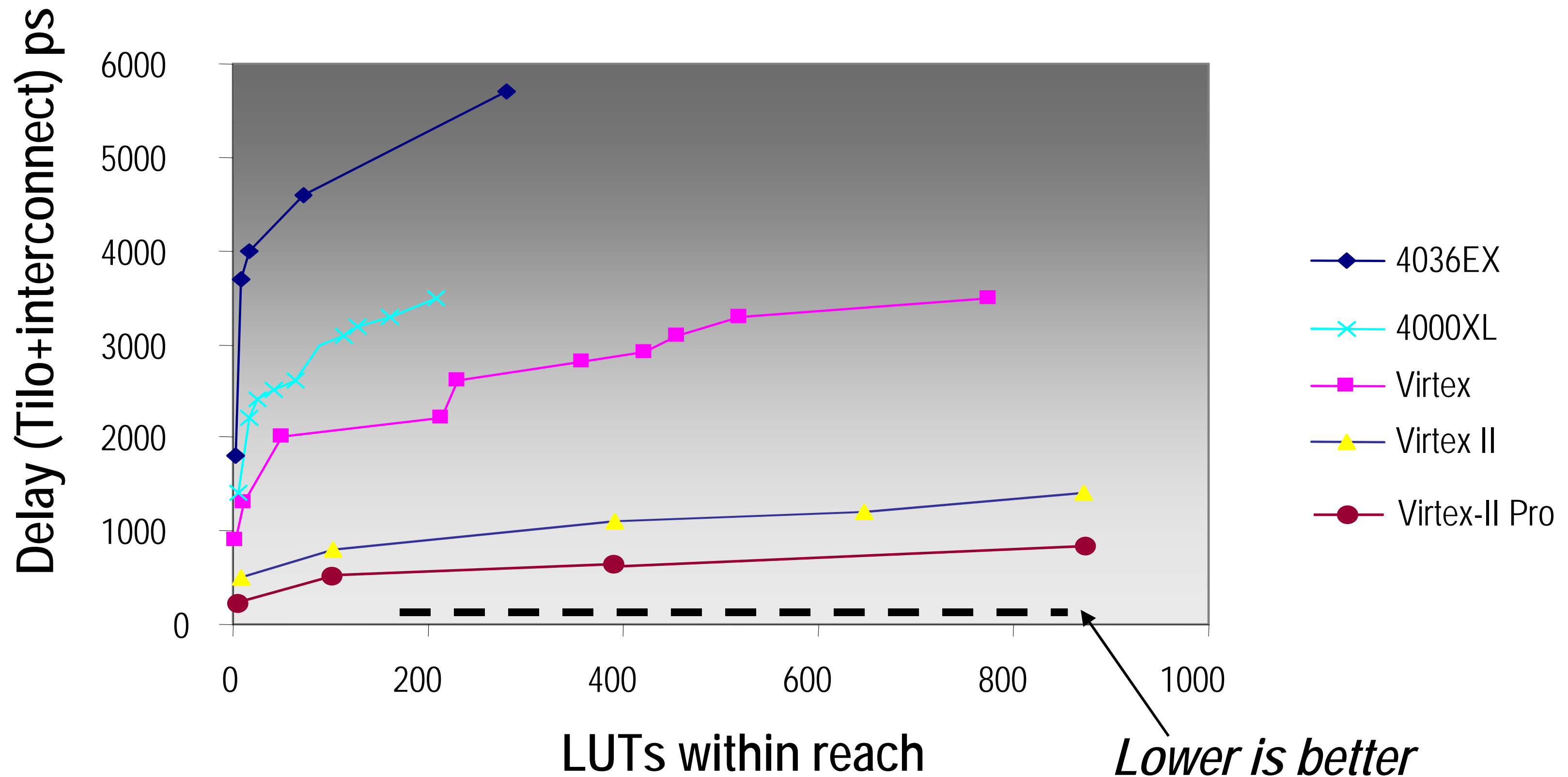
We were hit by fast-moving CMOS process technology, particularly multiple metal layers.

Trends



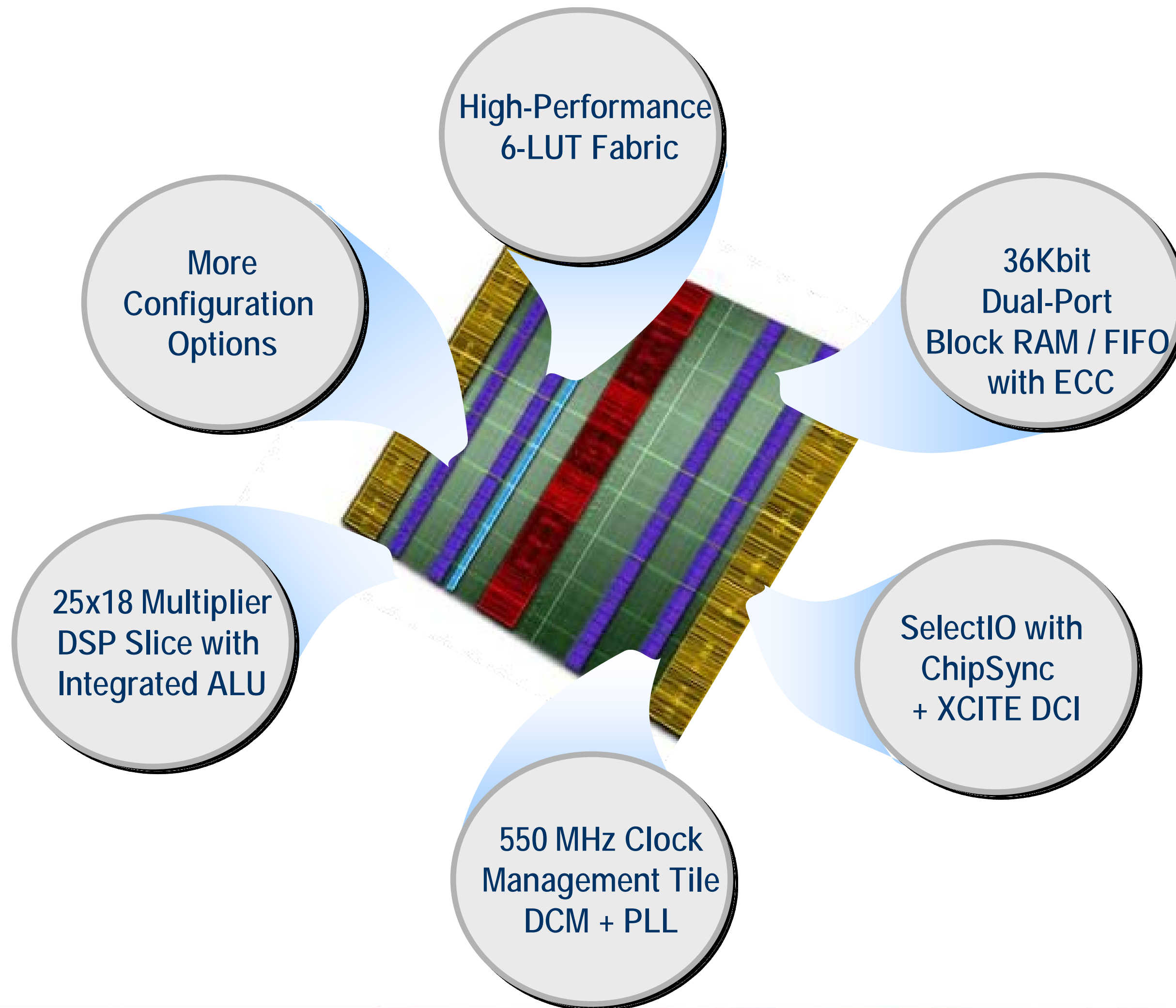
Courtesy and Copyright of UMC

Interconnect and ease of use



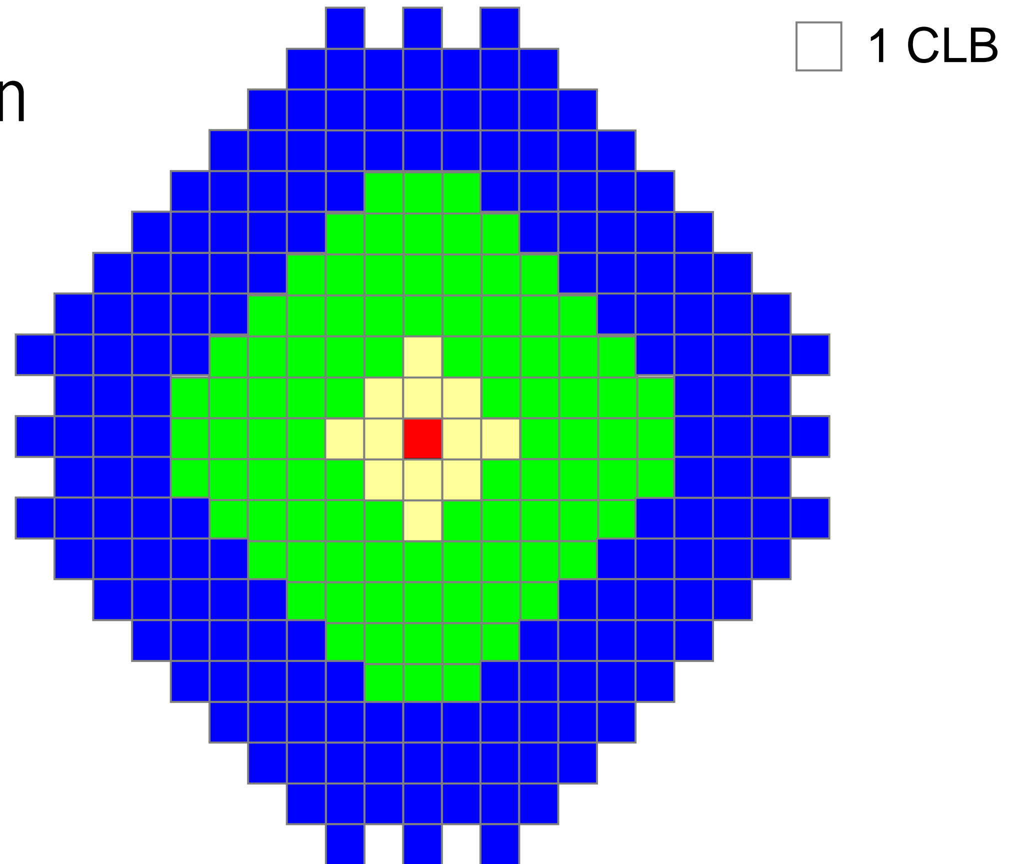
Longer wire reach gave dramatic improvement for ease of design

State-of-the-Art 65nm FPGA



Predictable Interconnect

Symmetric routing pattern reaches more CLBs with fewer hops

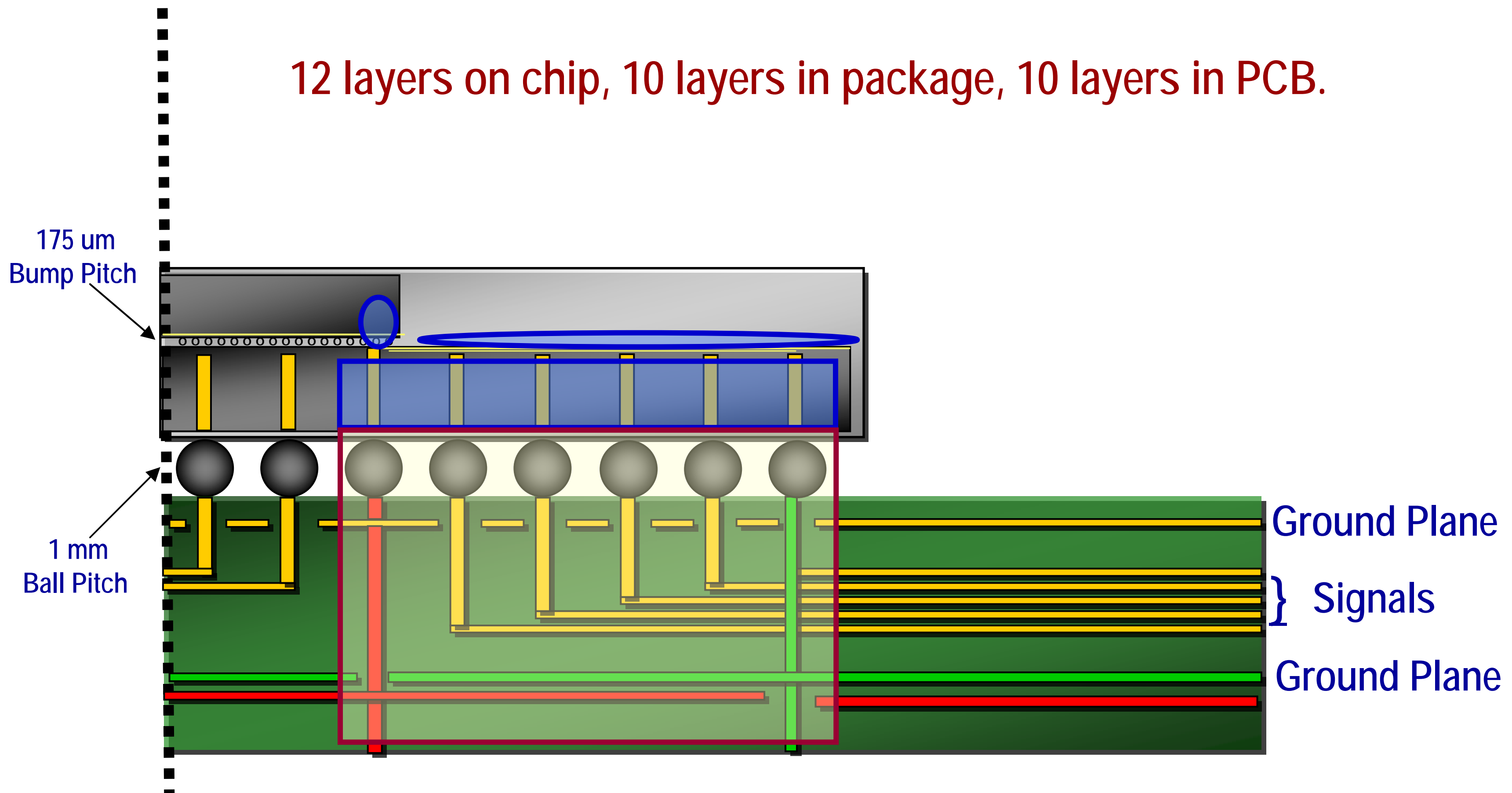


- Fast Connect
- 1 Hop
- 2 Hops
- 3 Hops

Dramatically increases design performance

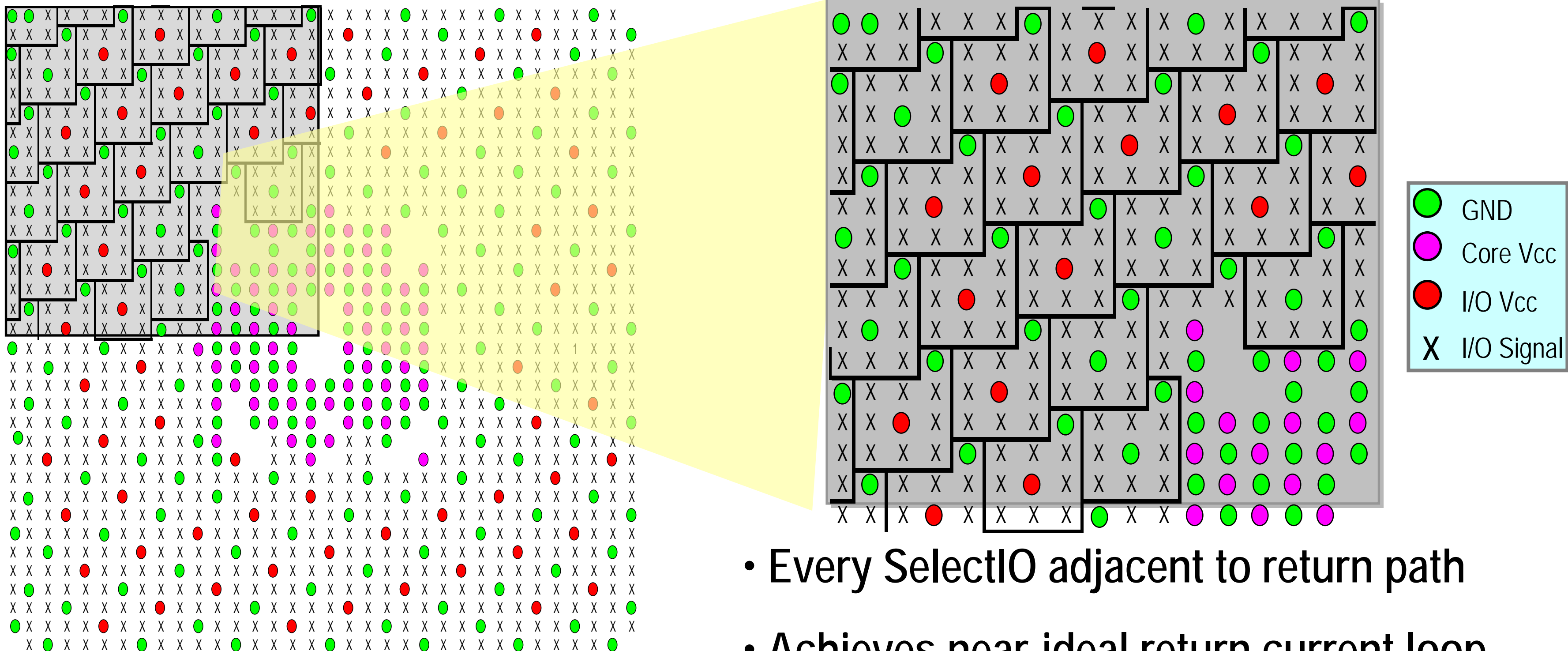
Layers of Interconnect

12 layers on chip, 10 layers in package, 10 layers in PCB.



82% of noise is determined by the pin-out and found in package balls and PCB vias

Sparse Chevron Pin-out Pattern

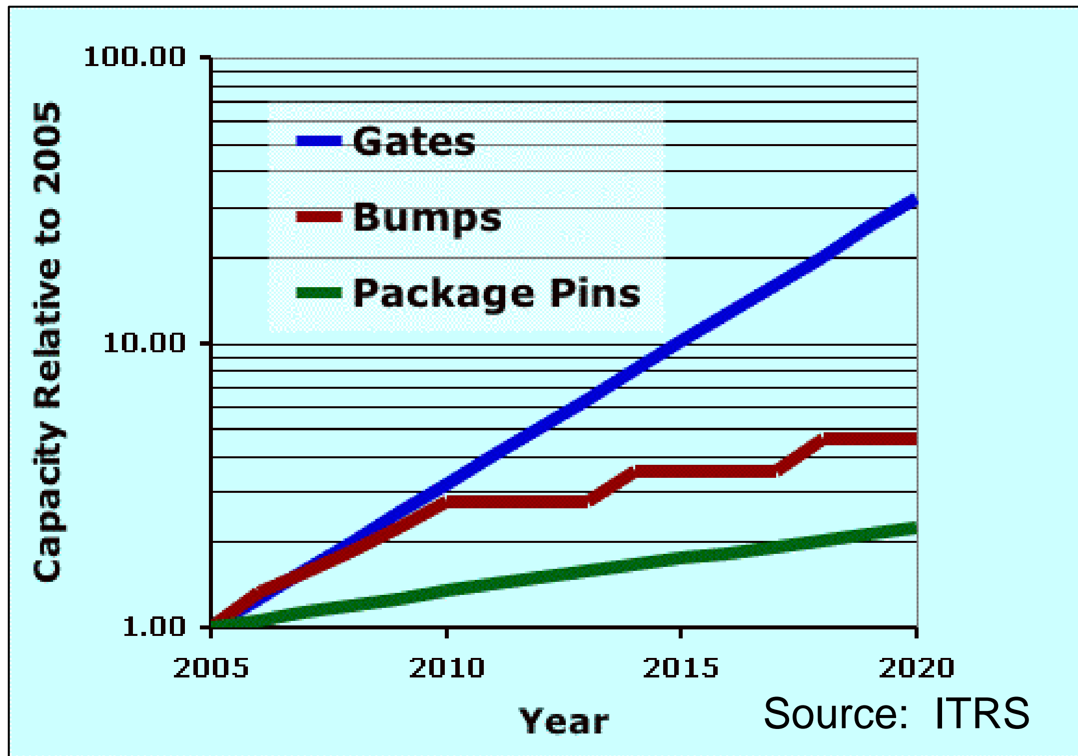


- Every SelectIO adjacent to return path
- Achieves near-ideal return current loop

IO pins in a regular array of return path pins

Off chip interconnect

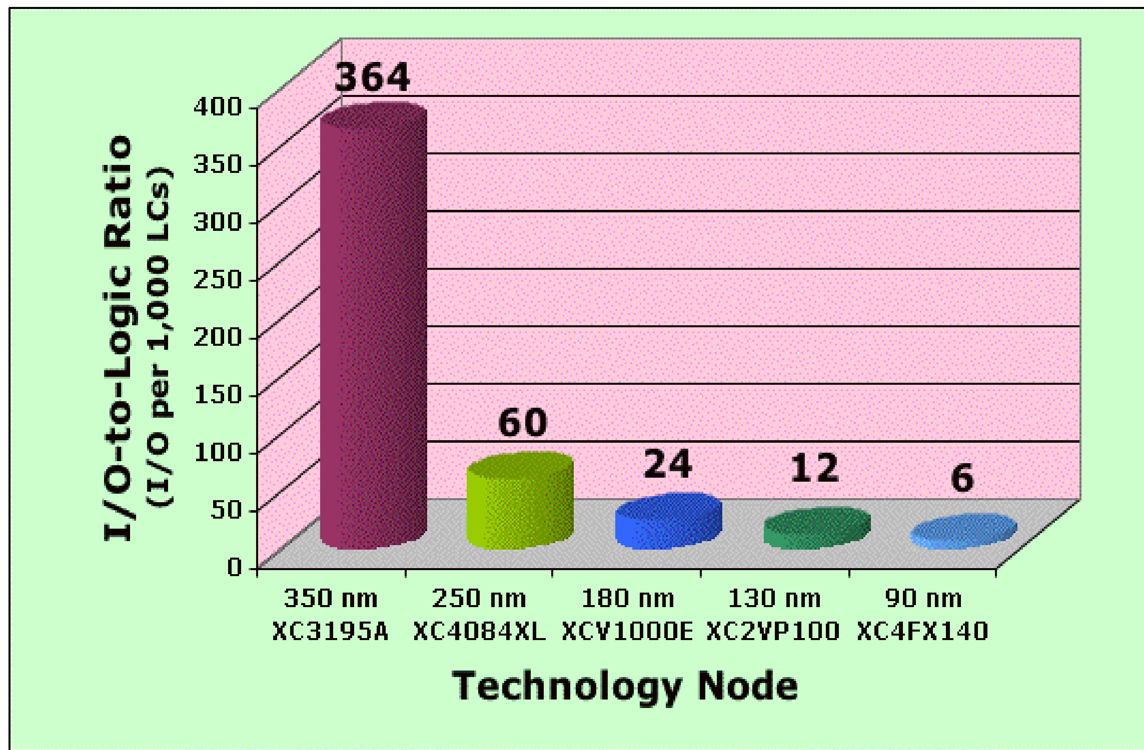
- Growing gap between number of logic gates and I/O
- Technology scaling favors logic density



15x drop in I/O-to-logic ratio by 2020

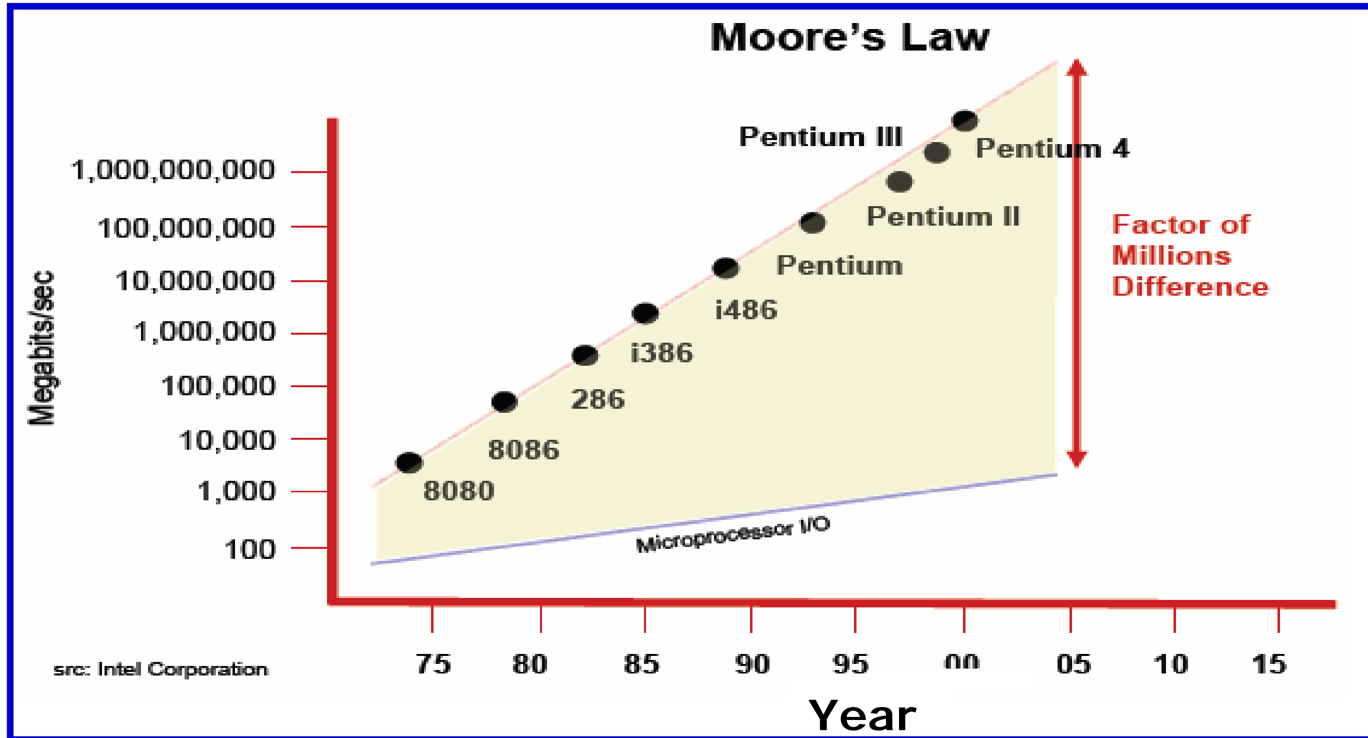
I/O to Logic Ratio

Comparison of number of I/O per 1000 logic cell in the largest FPGA in each family



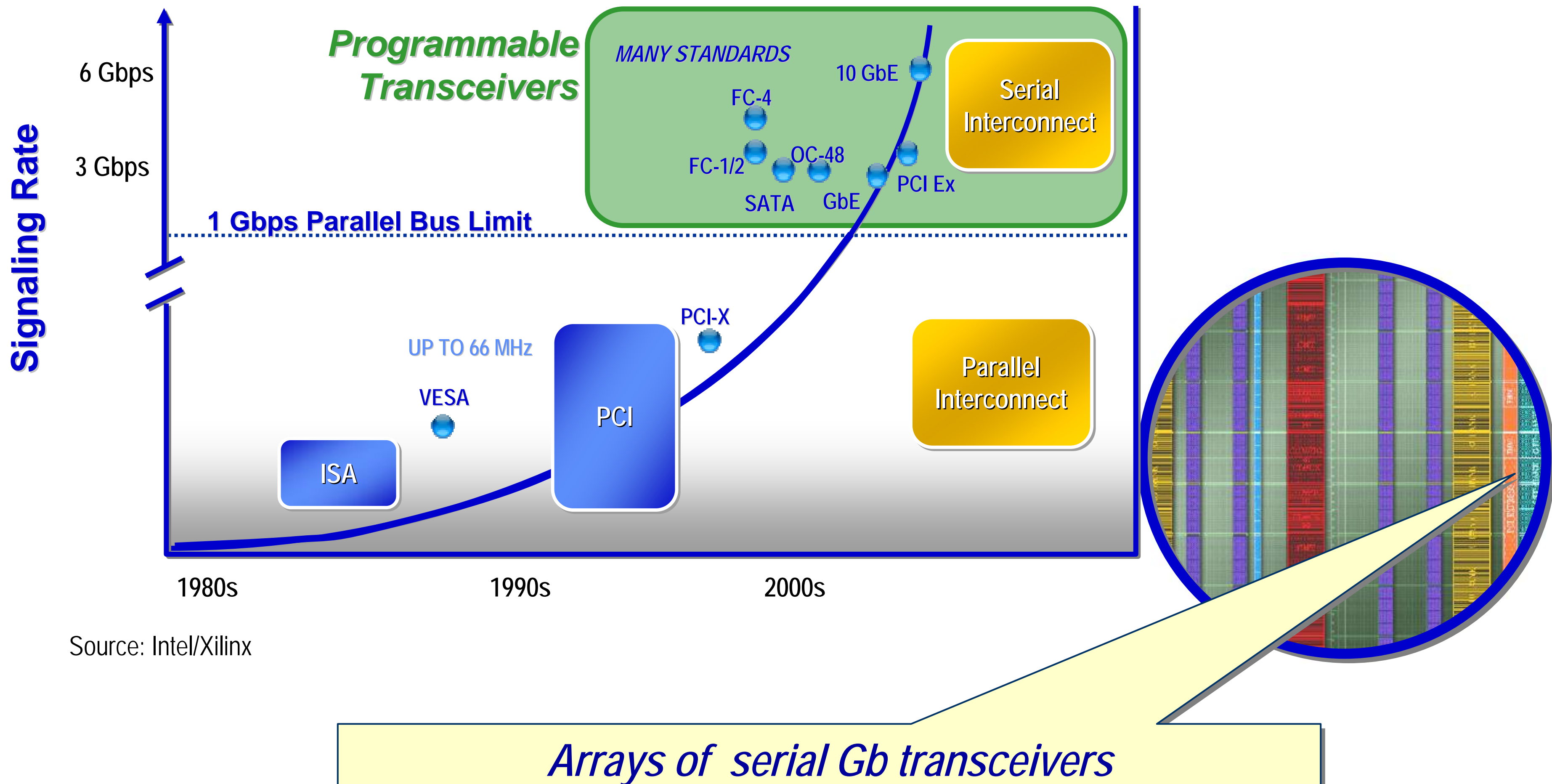
~60x decrease in I/O-logic ratio

Logic to I/O Gap in Microprocessors



Microprocessor's chip area is dominated by cache memory to overcome the lack of I/O bandwidth

The Move to Serial Connectivity



Die-Stacking Landscape

(Connection Density, Number of Device Layers)

10^2-10^3/cm^2
4+ device layers

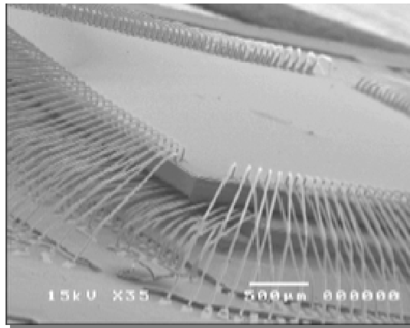
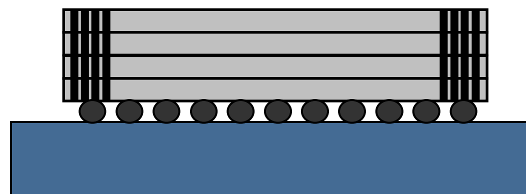


Photo: Amkor

Chip-Stacking (wire-bonding)

10^4 -10^6/cm^2
4+ device layers



Chip-Stacking (Through Silicon Vias)

>10^6/cm^2
3-4 device layers

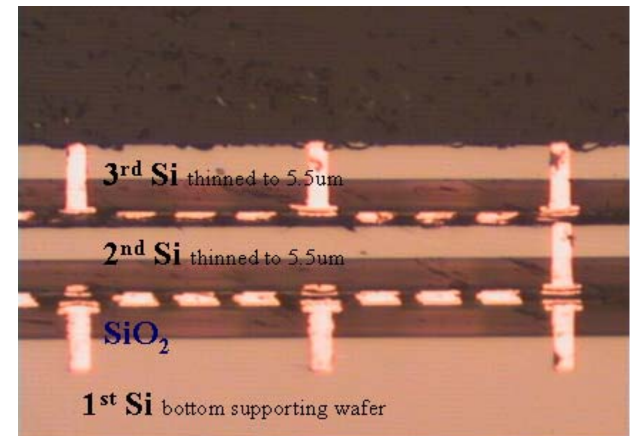


Photo: Tezzaron

Monolithic 3-D integration

Imagine…

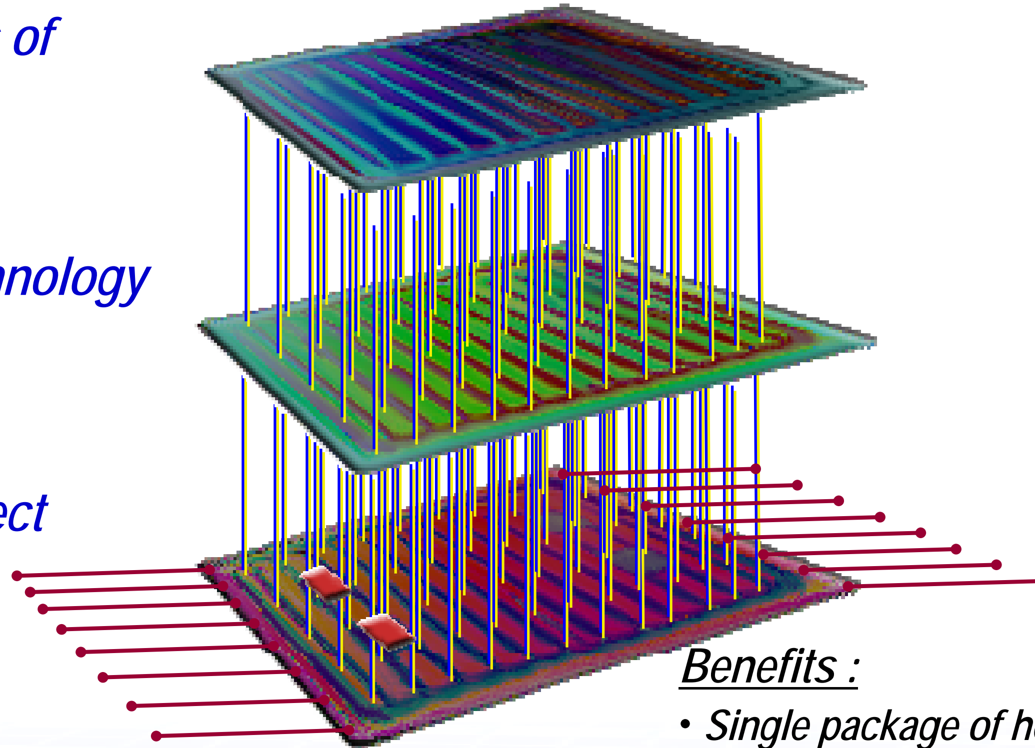
Specialized Layers of
DSP fabric,
Memory fabric,
FPGA fabric, …

Optimized for Technology
130nm,
90nm,
32nm, …

With 3D Interconnect

And at its heart…

An FPGA SoC with:
- *Embedded Processing*
- *Embedded DSP*
- *High-speed Serial Connectivity*
- *Reprogrammable FPGA Logic Fabric as the Base*

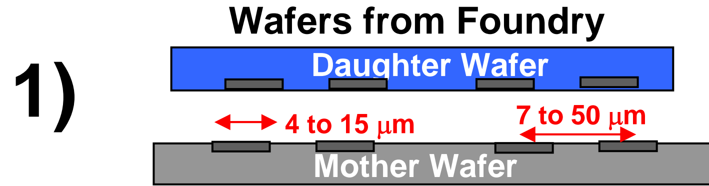


Benefits :
- *Single package of heterogeneous die*
- *Multiple configurations of standard products*
- *Lower cost*
- *Lower power*
- *Ultimate customization*
- *Ultimate flexibility*

Wafer Bonding

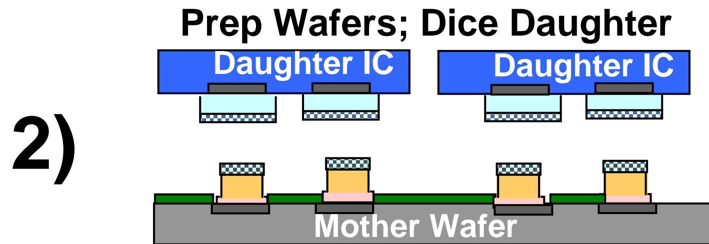
1)

Wafers from Foundry



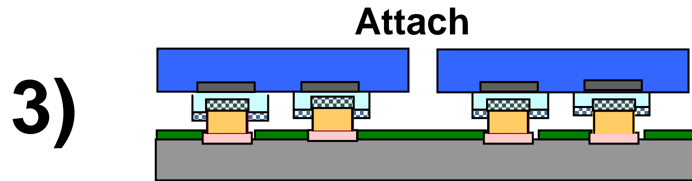
2)

Prep Wafers; Dice Daughter



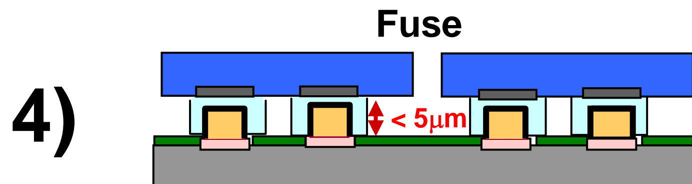
3)

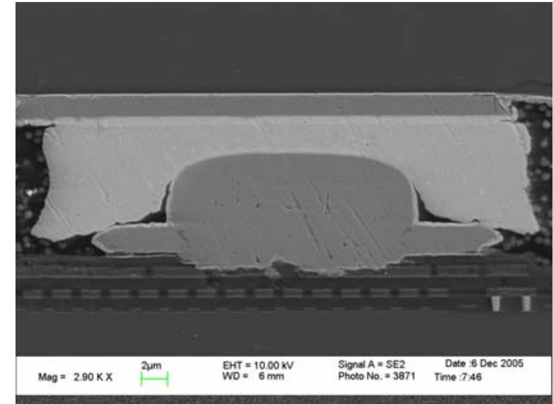
Attach

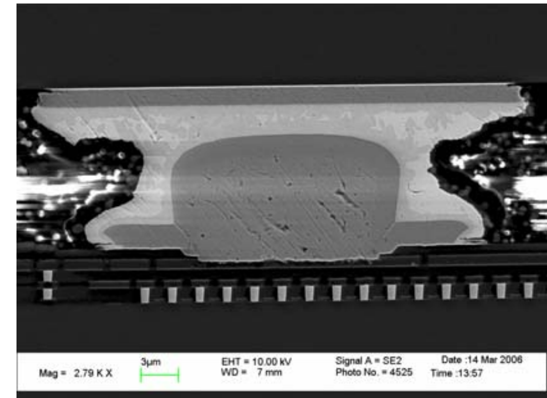


4)

Fuse

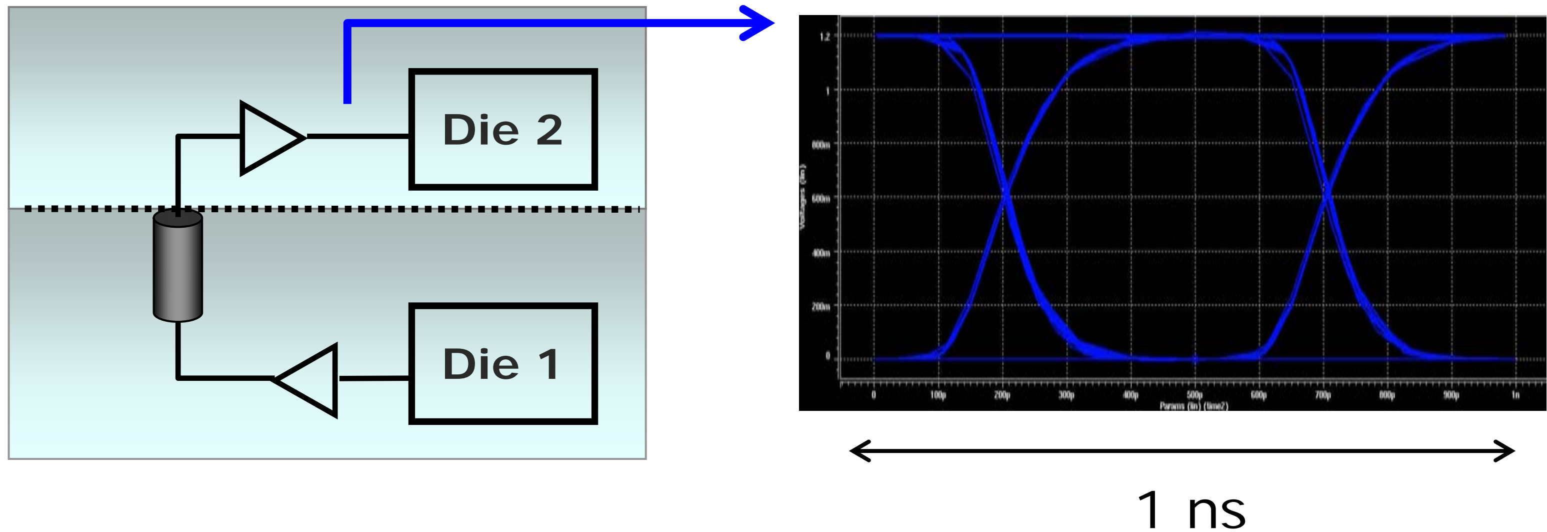






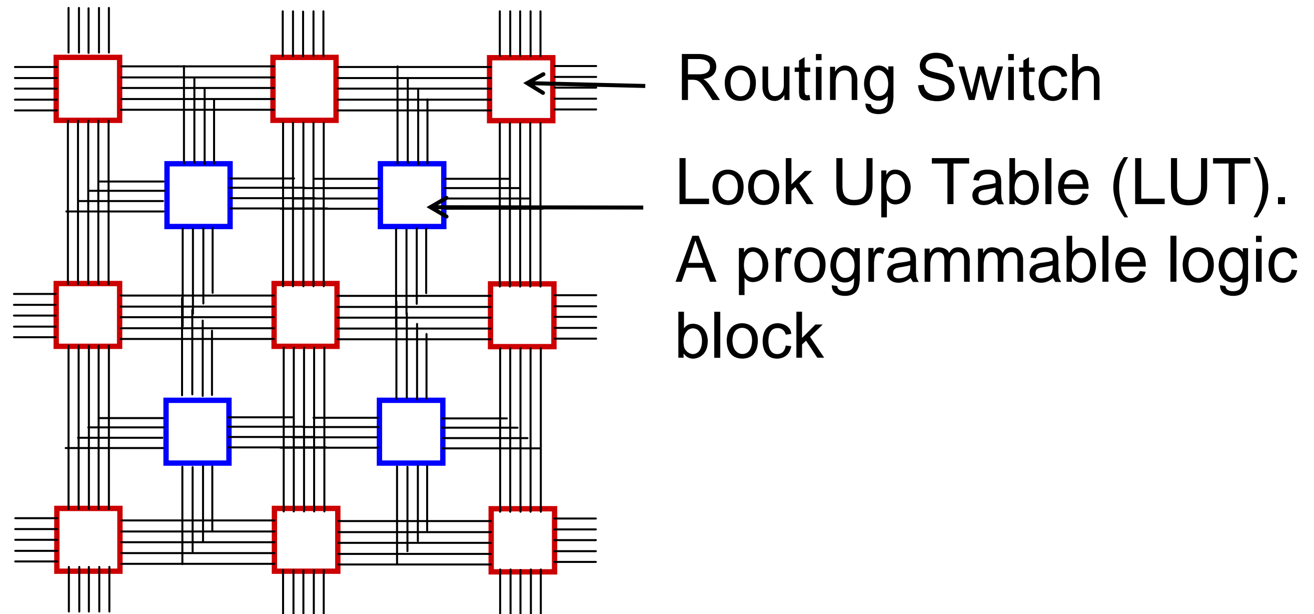
Source: Cubic Wafer

Inter-Die Connection Performance

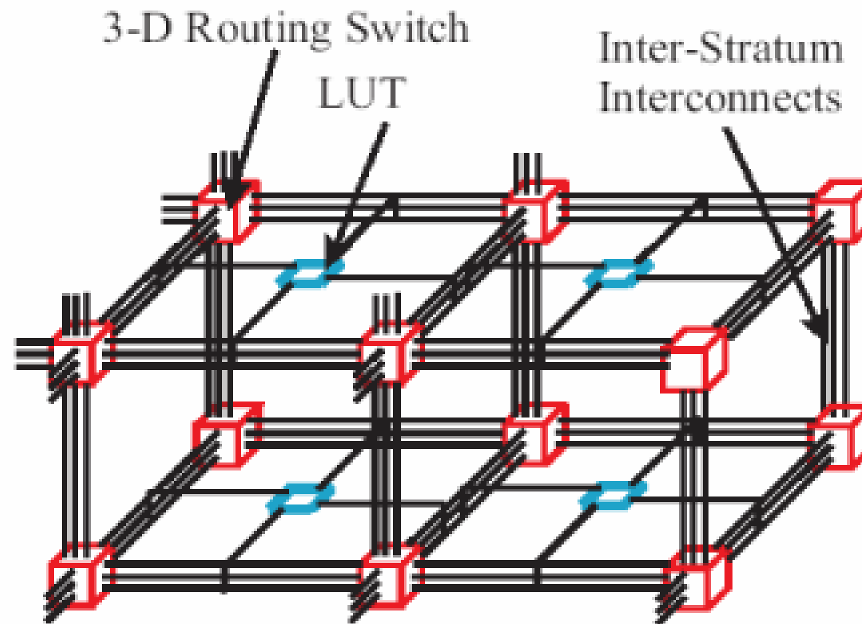


- 100X less dynamic power than conventional single-ended I/O
- Link performance ~ 1 Gigabit/sec

2-D FPGA Fabric



3-D FPGA Fabric



A. Rahman, et al., IEEE TVLSI, 11(1), 2003

- Shorter wire-length and delay
- Higher logic density

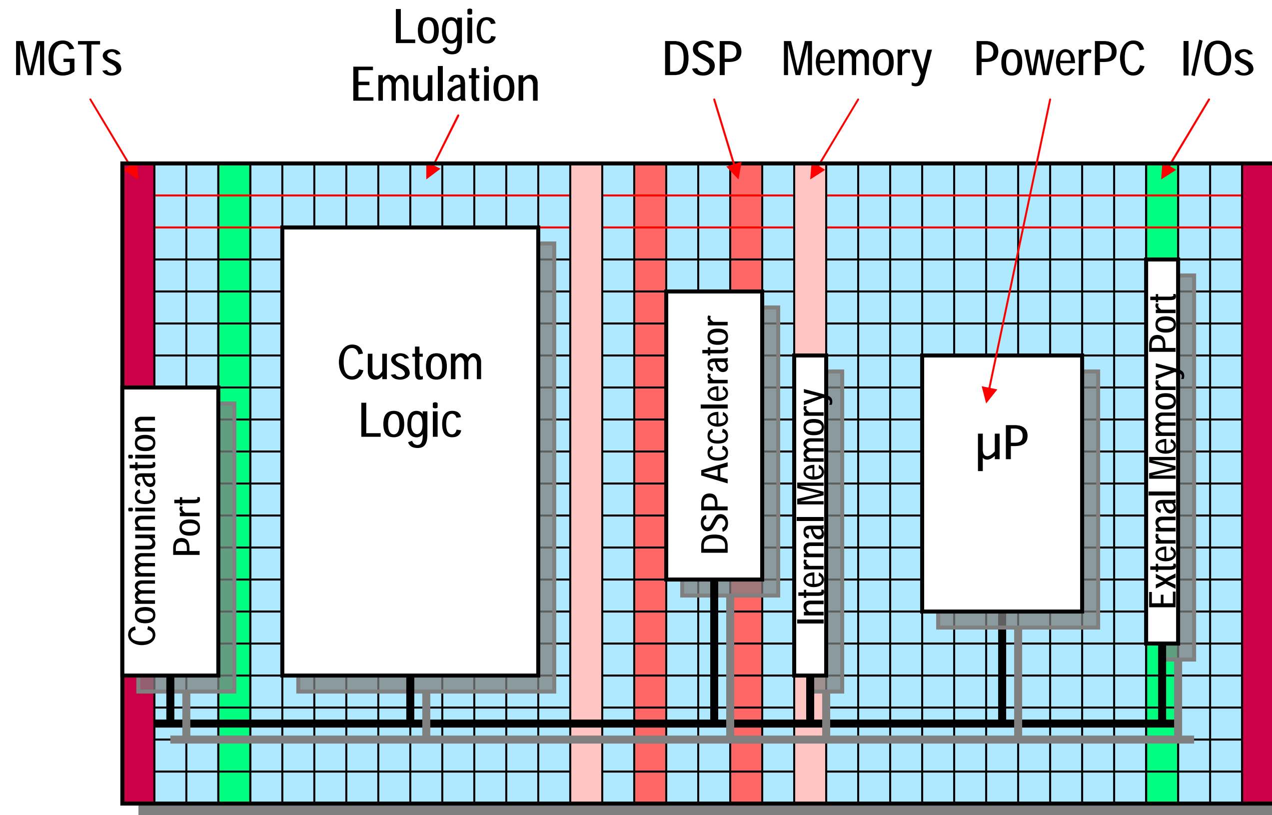
Performance Improvement

- Integration of RAM with FPGA by high bandwidth die-stacking
- 2 Terabit/sec bandwidth between FPGA and RAM

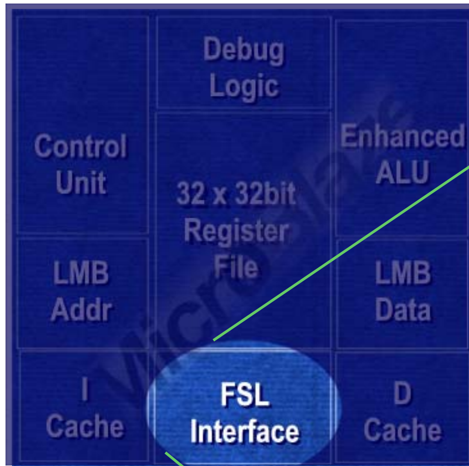
High Performance Applications	Projected Performance Improvement
Sparse Matrix/ Vector Multiply	4X-8X over 2.4GHz Pentium
Traffic Simulation	170X over 2.2GHz Opteron
Radiative Heat Transfer	20X over 1.7GHz Pentium
Molecular Dynamics	~20X over 3 GHz Processor

(Courtesy of Maya Gokhale, LANL)

The FPGA System



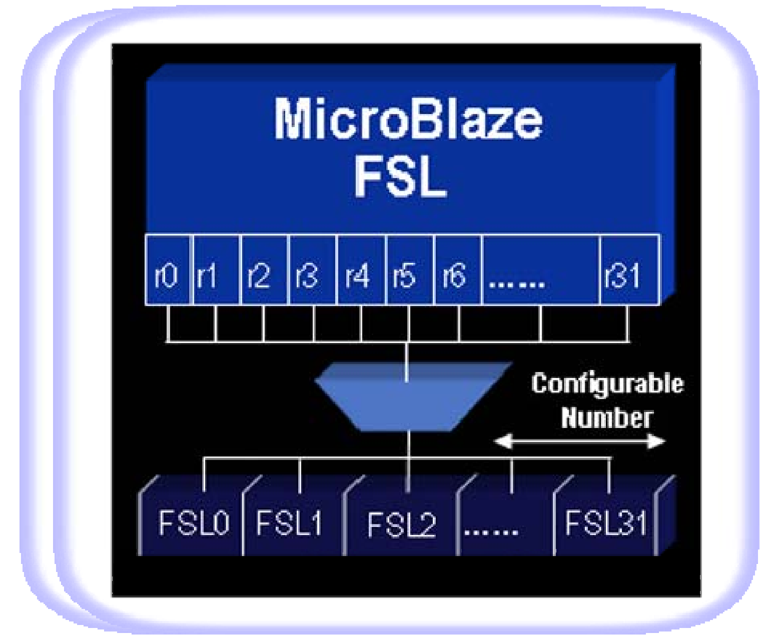
MicroBlaze's Flexible Acceleration Interface

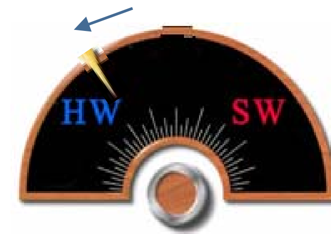


- FSL = *Fast Simplex Link*
- Eliminates bus signaling overhead
  - No address decode
  - No arbitration
  - No acknowledge cycles
- Simple instruction programming
- Flexible number master and slave FSL ports
  - Configurable depth FIFO in FSL
  - Input and output FSL port width is configurable as 8,16, or 32 bits.
- Dedicated MicroBlaze instruction
  - Get fromInputFSL M, toReg N
  - Put toOutputFSL M, fromReg N
  - Blocking and non-blocking support

Application-Specific Hardware Acceleration

- When the processor core begins to reach software task capacity, then Fabric Acceleration to the rescue
  - Use Fast Simplex Link (FSL) to interface to customer-defined accelerators
  - Enables dramatic improvements in performance



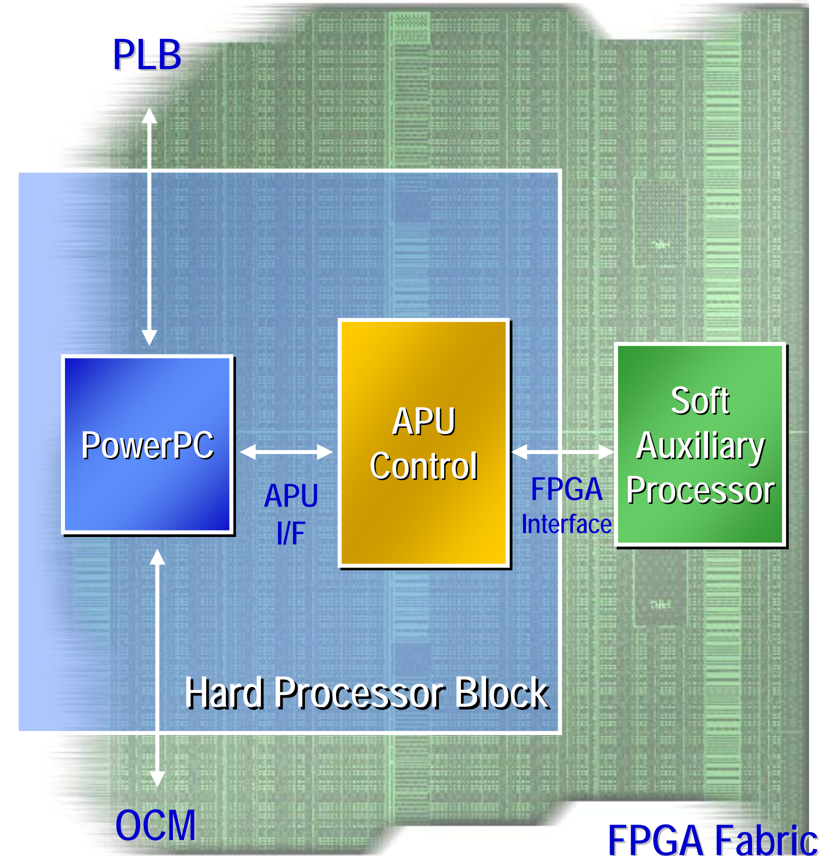


FPGA/processor

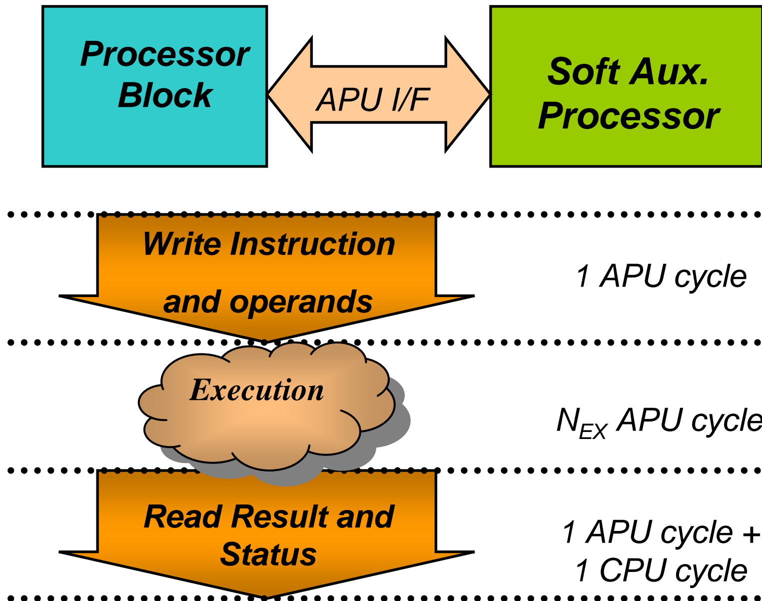
- CoreConnect Architecture
  - Processor Local Bus (PLB)
    - Ideal for bust transfers
      - Memory, High Speed Peripherals, Cache Interface
    - 32-bit address, 64-bit data
    - 2.1 GB/s Max BW @ 133 MHz
  - On-Chip Peripheral Bus (OPB)
    - Low speed peripherals
    - 32-bit address, 32-bit data
  - Device Control Register Bus (DCR)
    - For peripheral setup and control
- On Chip Memory Interface (OCM)
  - 4 Processor Cycle Latency
  - Lowest Latency and good data rate
  - Not a bus interface

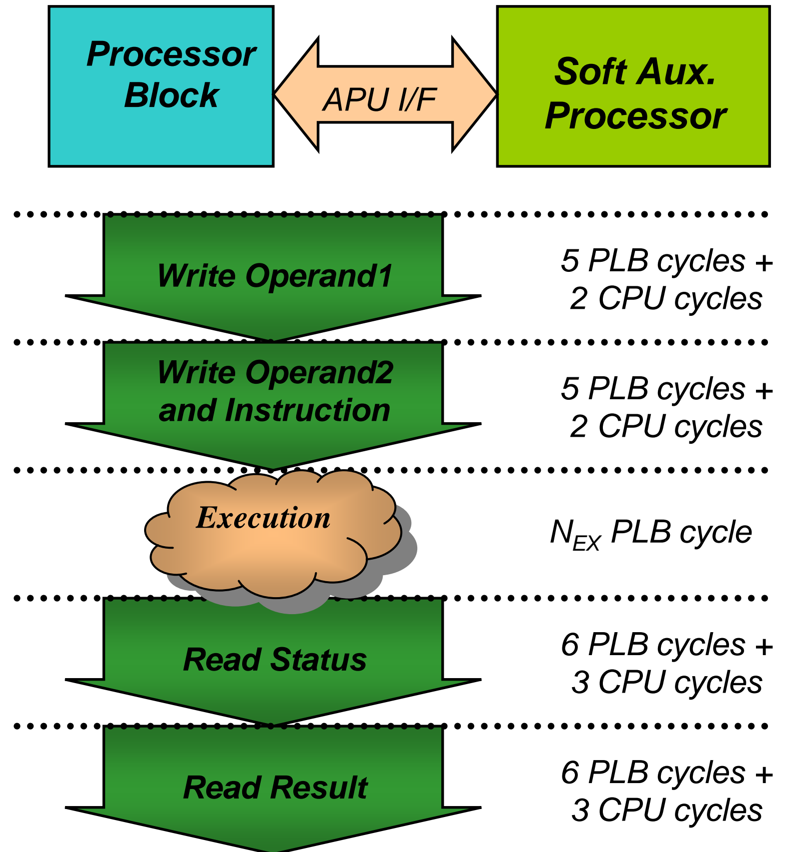
Accelerate Performance Beyond the Core

- Extends PPC 405 Instruction Set
  - Floating point support
  - User Defined Instructions
- Offloads CPU intensive operations
  - Matrix calculations
    - Video processing
  - Floating point mathematics
    - 3D data processing
- Direct interface to HW accelerators
  - High Bandwidth
  - Low Latency
- Reduce number of bus cycles by factor of 10X
- Increase performance by over 20X

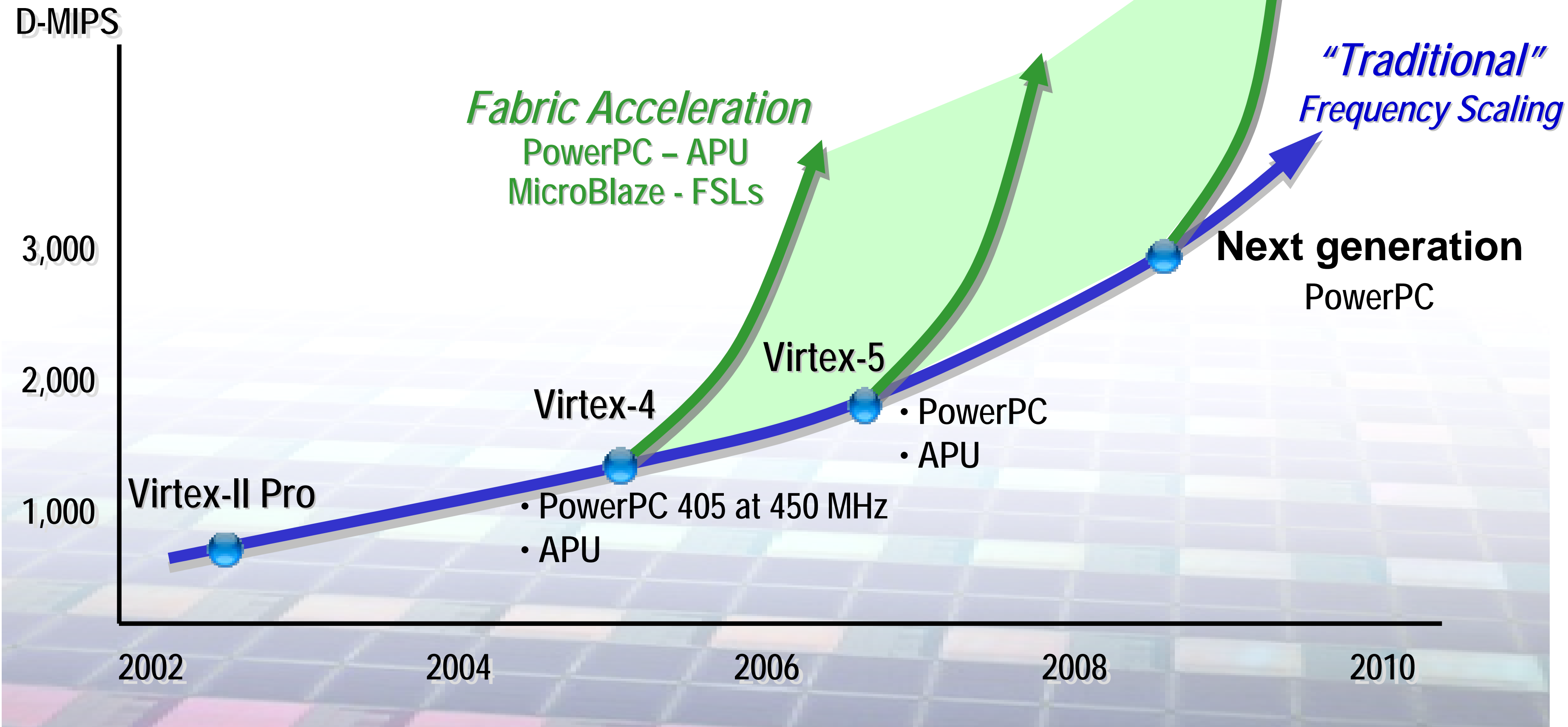


Comparison with Traditional Bus-based

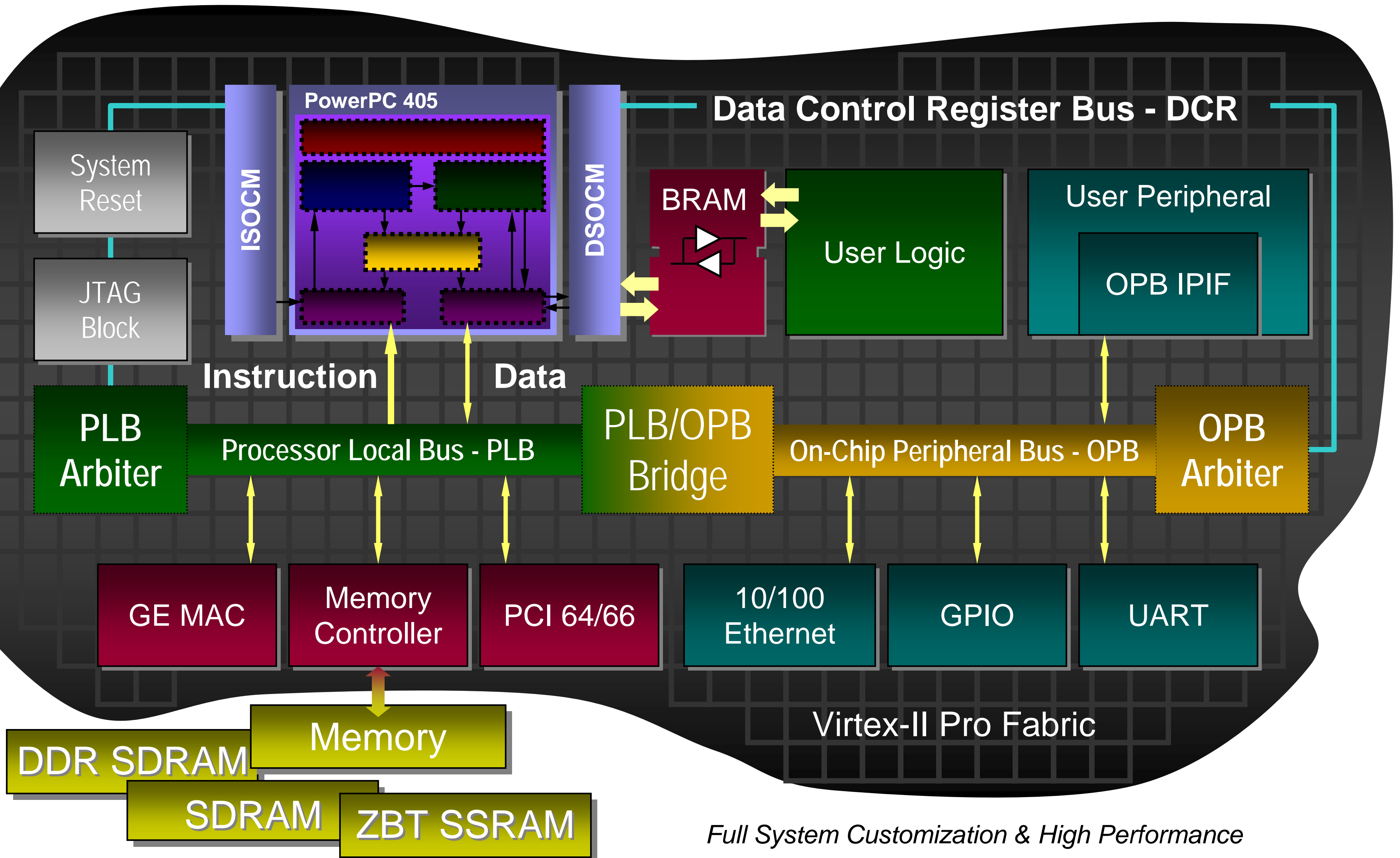




Processor Performance and Fabric Acceleration

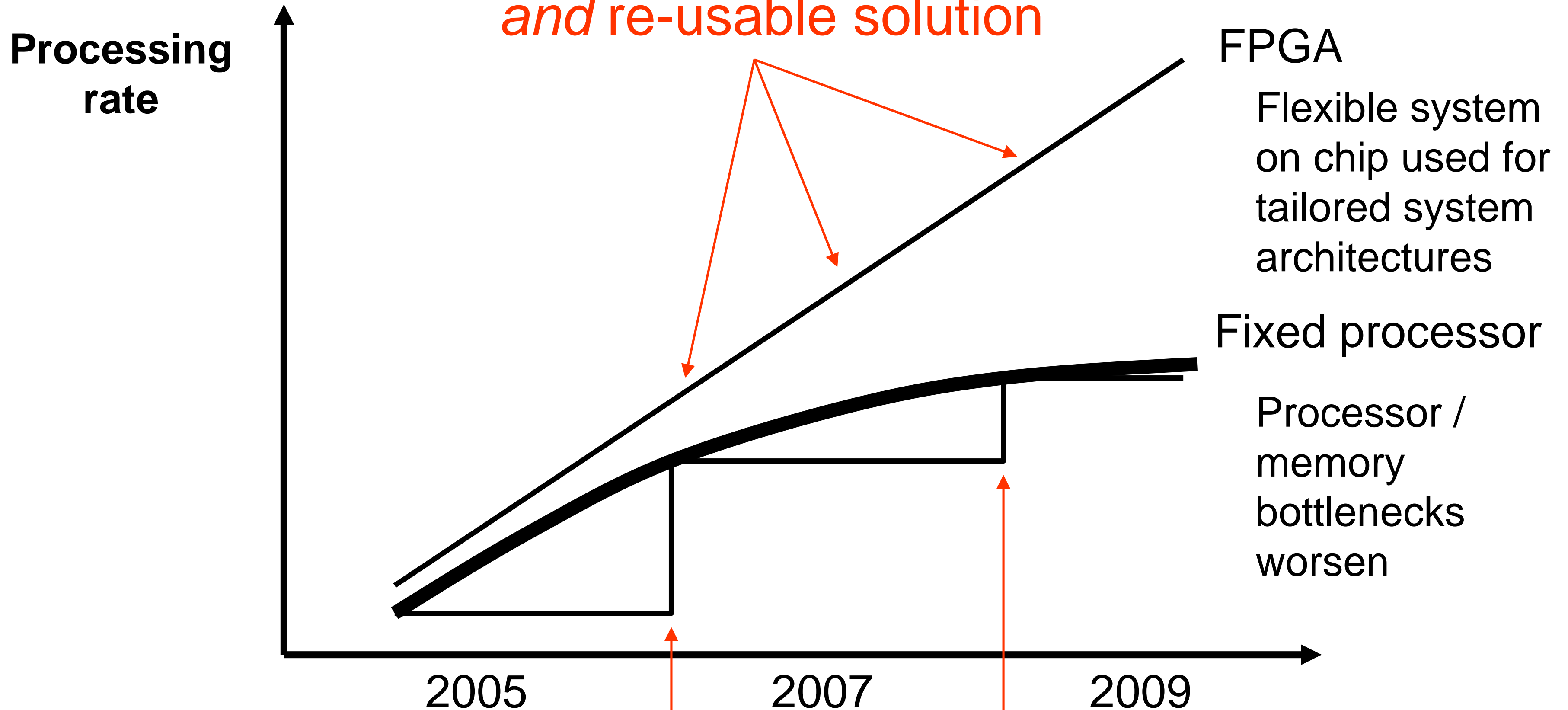


PowerPC Architecture



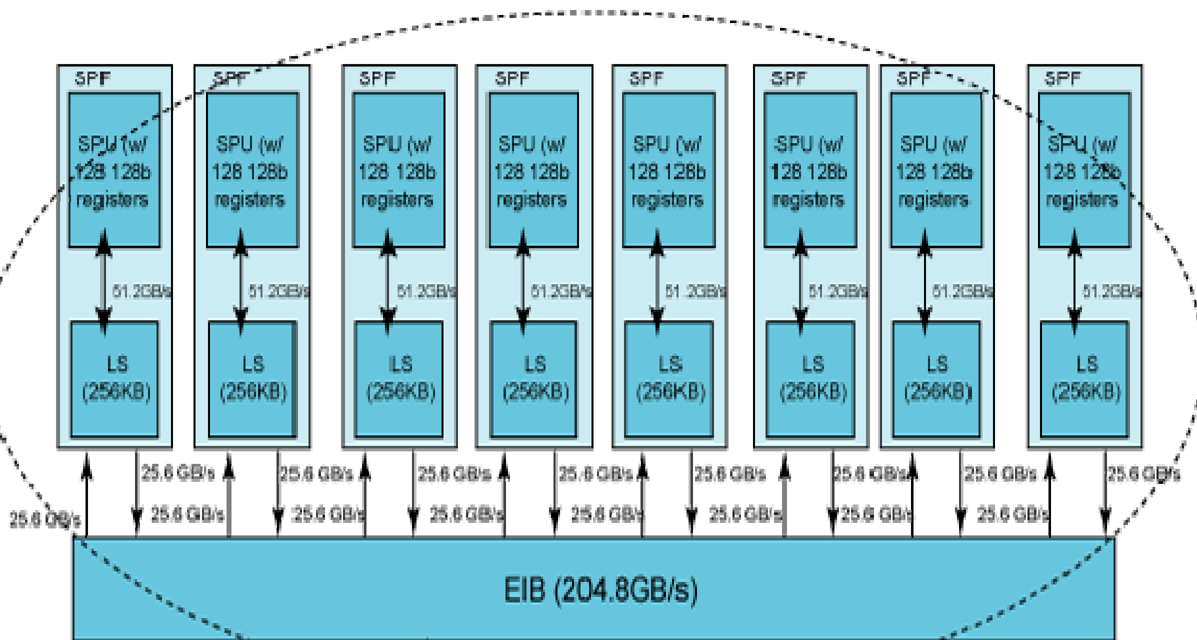
Full System Customization & High Performance

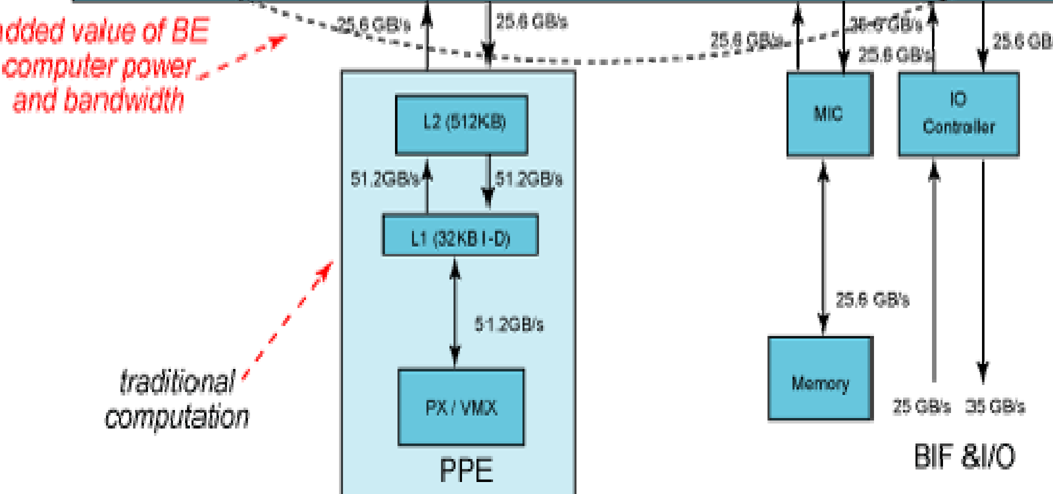
Scalability



No re-use, architecture dependent

Configurable Interconnect





Cell Processor

- 85GB External Bandwidth
- 205GB internal bus
- 400GB register file bandwidth
- 400GB internal memory(cache)
- 60-130 Watts

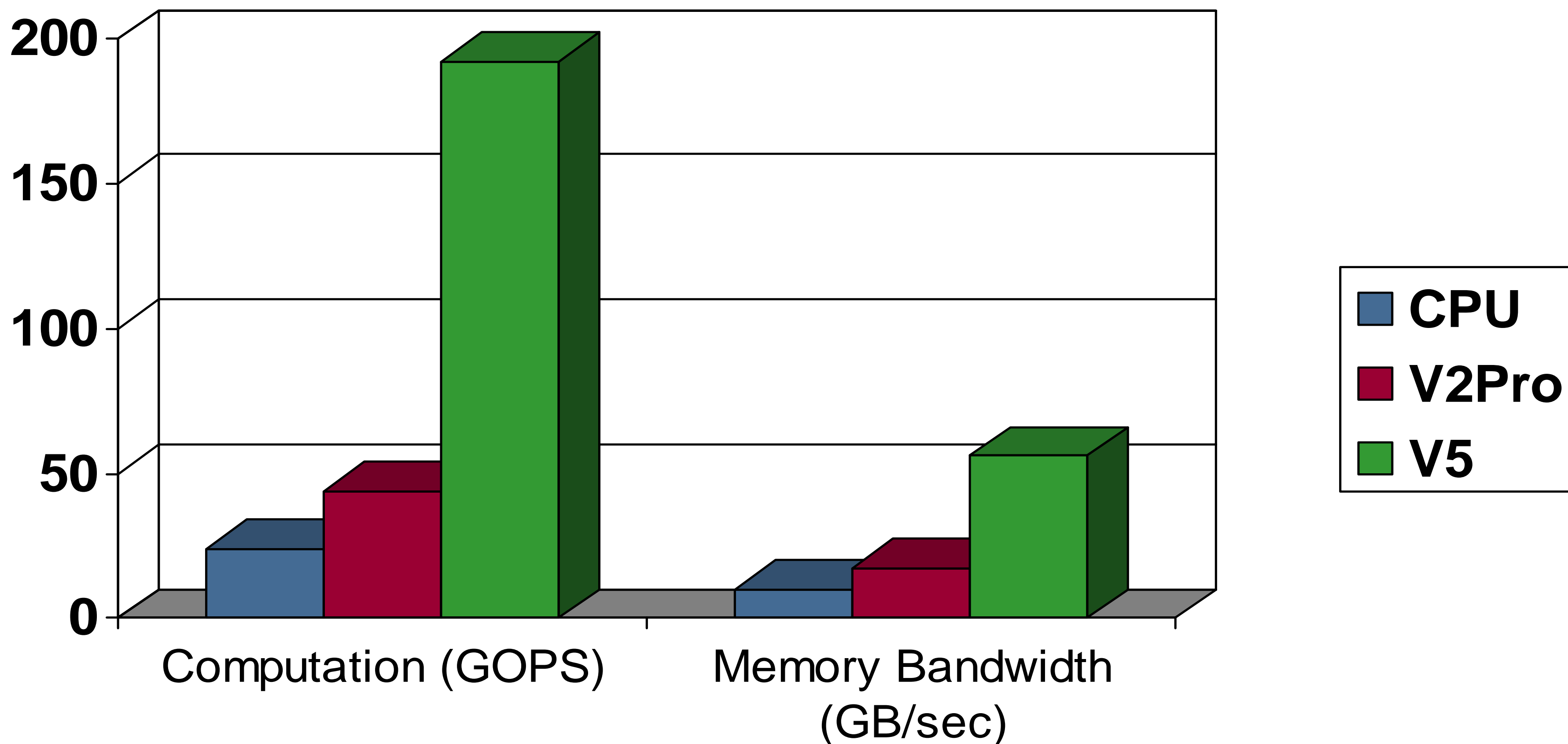
FPGA

- 80GB External Bandwidth
- 2TB internal memory(BRAM)
- 2.5TB internal embedded mem
- >10TB ALU<->register<->ALU
- 10-20 watts

You must customize your program to make use of the busses of the Cell processor

The FPGA connectivity can be customized to fit your program

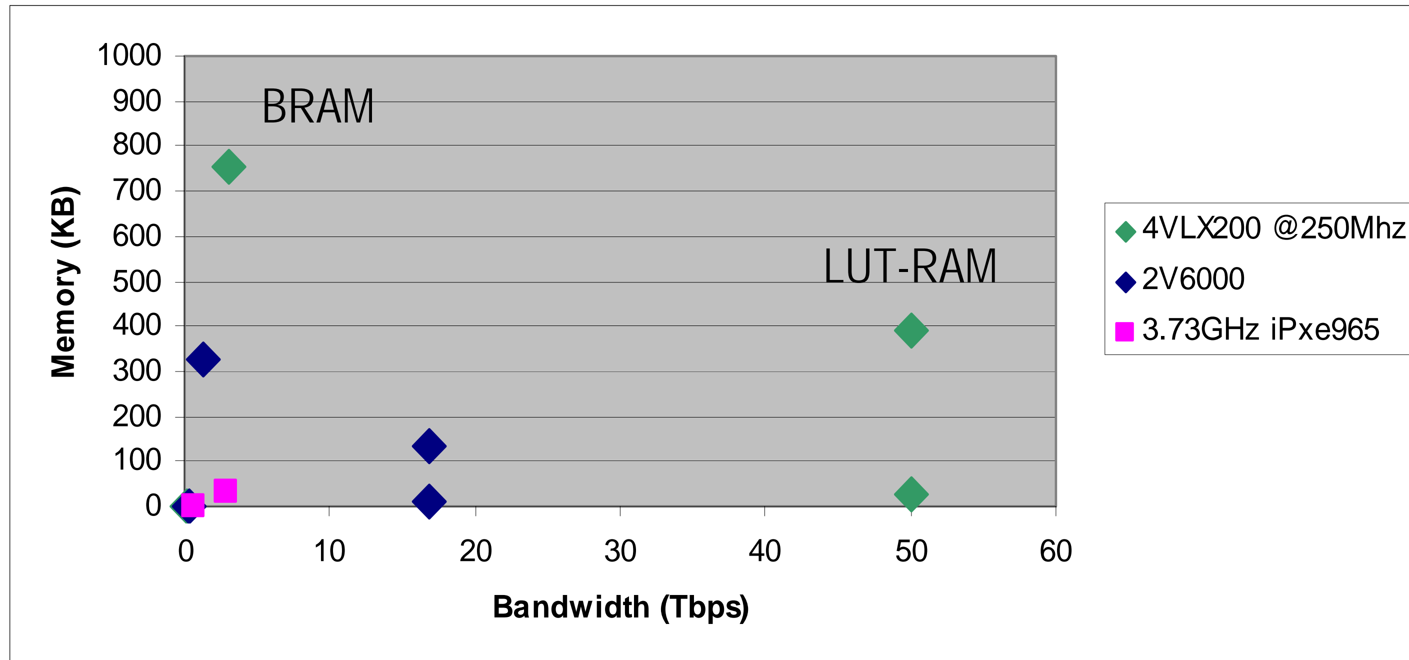
Flexibility



Woodcrest : 24 GFlops @3Ghz
Virtex-5 : 60 GFlops (70 GFlops-SP) @ 350Mhz

Woodcrest = 80 Watt
Virtex-5 = 10 Watt

Internal Memory Bandwidth



Conclusions

- It is all about connectivity
- Important aspects
  - Off-chip interconnect goes serial
    - Latency, power
  - On-chip interconnect has to be
    - Scalable
    - Hierarchical
    - Flexible
    - Easy to use
  - System interconnect heterogeneous and tailored to application