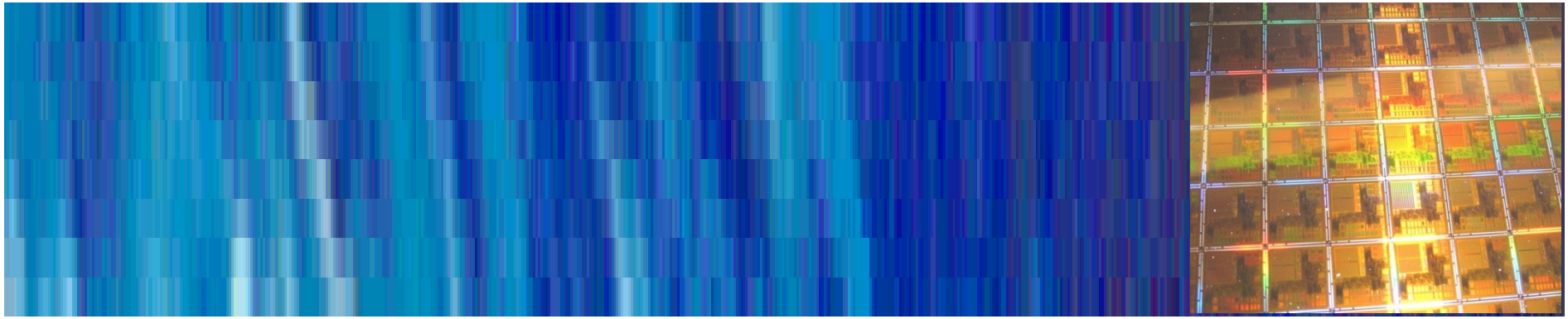
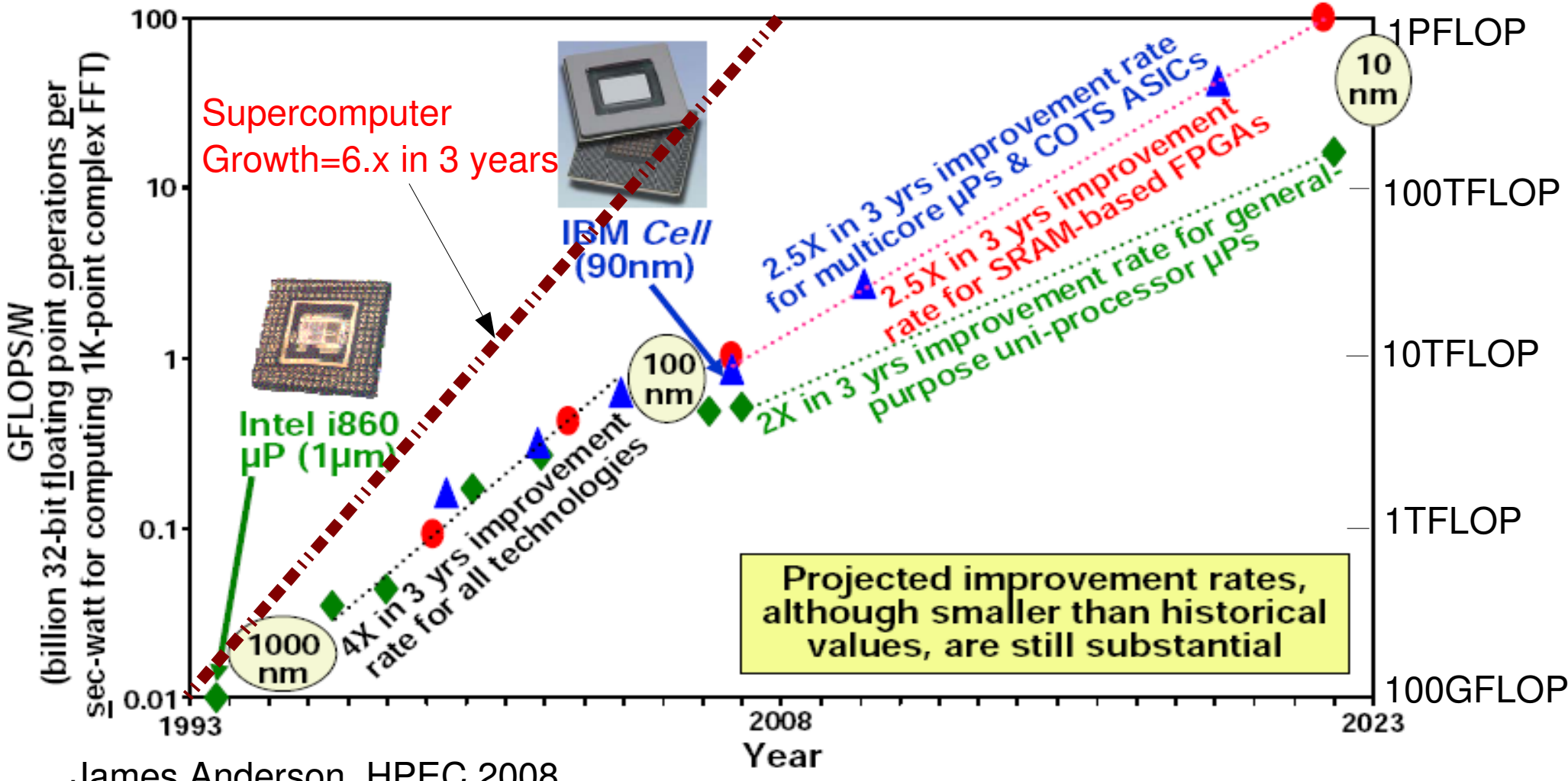


A Manycore Coprocessor Architecture for Heterogeneous Computing

Andreas Olofsson
andreas@adapteva.com

ADAPTEVA

Houston, we have a problem..

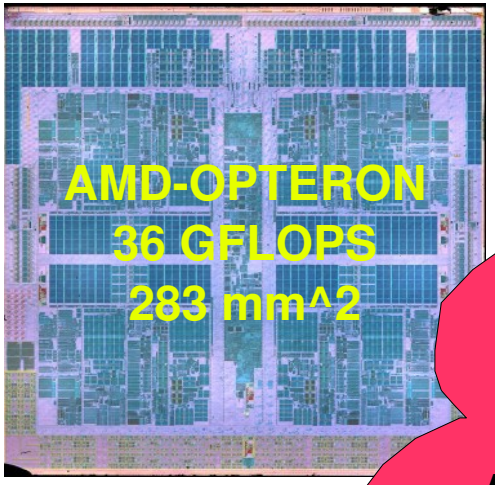


James Anderson, HPEC 2008

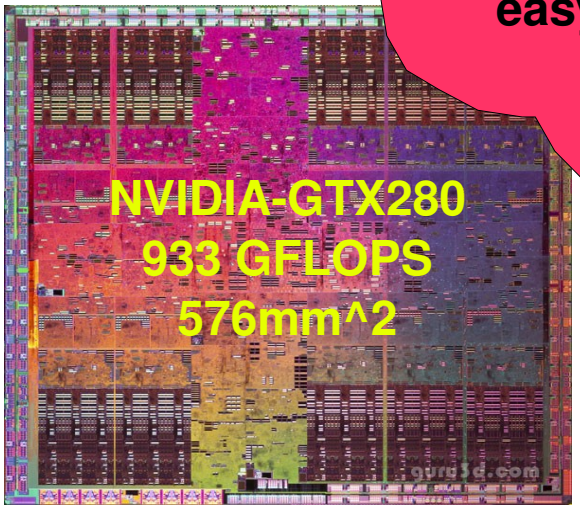
Current Path To Exaflop Supercomputing

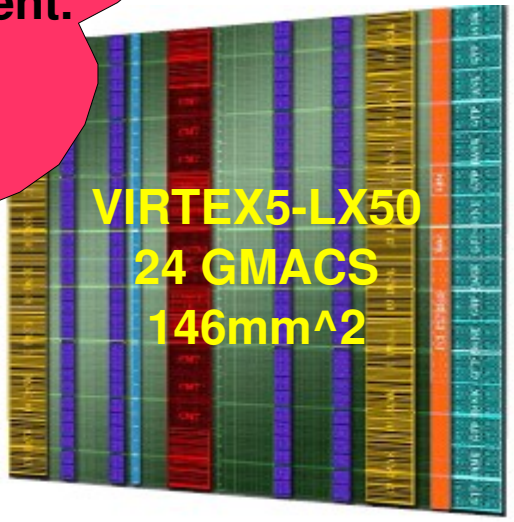
- **2023 Projection:**
  - 10 EXAFLOP
  - 100 GFLOP/Watt
  - 10 GWatt System
- **2023 Alternative:**
  - Change programming model
  - 10 EXAFLOP
  - 1-10 TFLOP/Watt
  - 0.1 – 1 GWatt

Ease of use → Area → Power



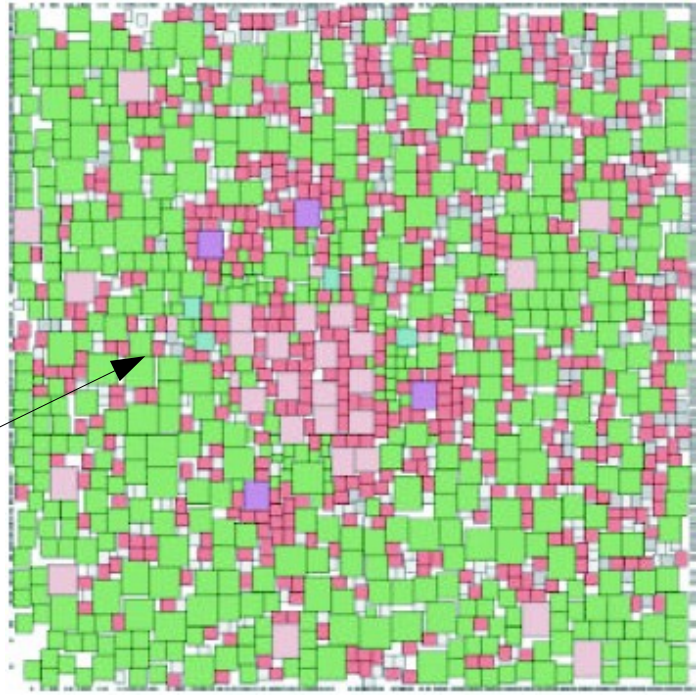
The quest for the holy grail :
An architecture that is flexible,
easy to use, and power efficient.





Stratix IV FPGA Example(40nm)

- 820K Logic Elements
- 48 high speed transceivers at 8.5 GB/s
- 23 MB SRAM
- 1288 hard macro multipliers(18bit)
- Up to 264 LVDS pairs

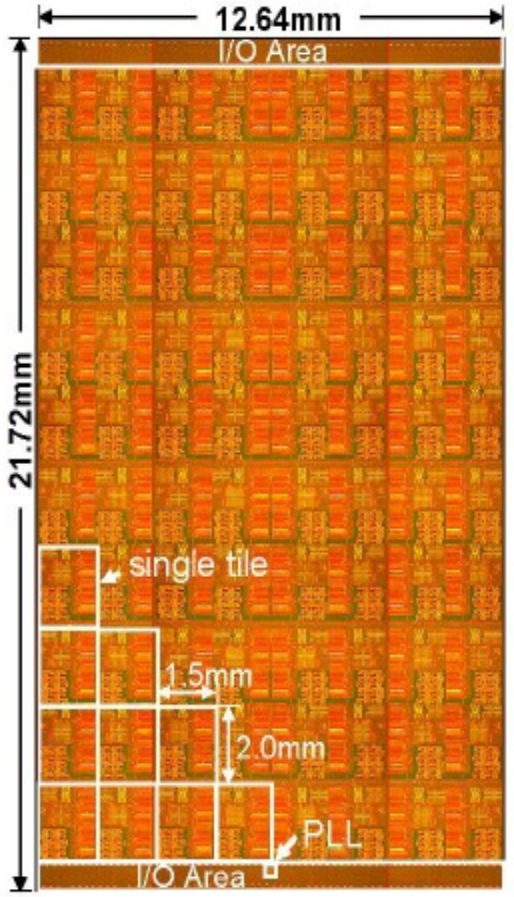


But, even FPGAs have warts
(blue and purple squares do useful work!)

Source: Rose, et al FPGA 2003

Intel *Teraflops* Effort (65nm)

- **Core Details:**
  - VLIW Machine
  - Dual FMADD
  - 2 Kbytes DMEM
  - 4 Kbytes IMEM
- **Interconnect Details:**
  - 2D Mesh NOC
  - 20 Gbyte/link BW
- **Performance:**
  - 5-25 GFLOPS/WATT



Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm^2
Tile area	3mm^2
Package	1248 pin LGA, 14 layers, 343 signal pins

Figure 5.2.7: Full-Chip and tile micrograph and characteristics.

Usage Programming Model

FPGA:
- Interfacing
- Data Flow
- Bit Level processing

Soft Coprocessor

Coprocessor:
- Great at math!
- Limited flexibility
- Simple programming
- High Efficiency

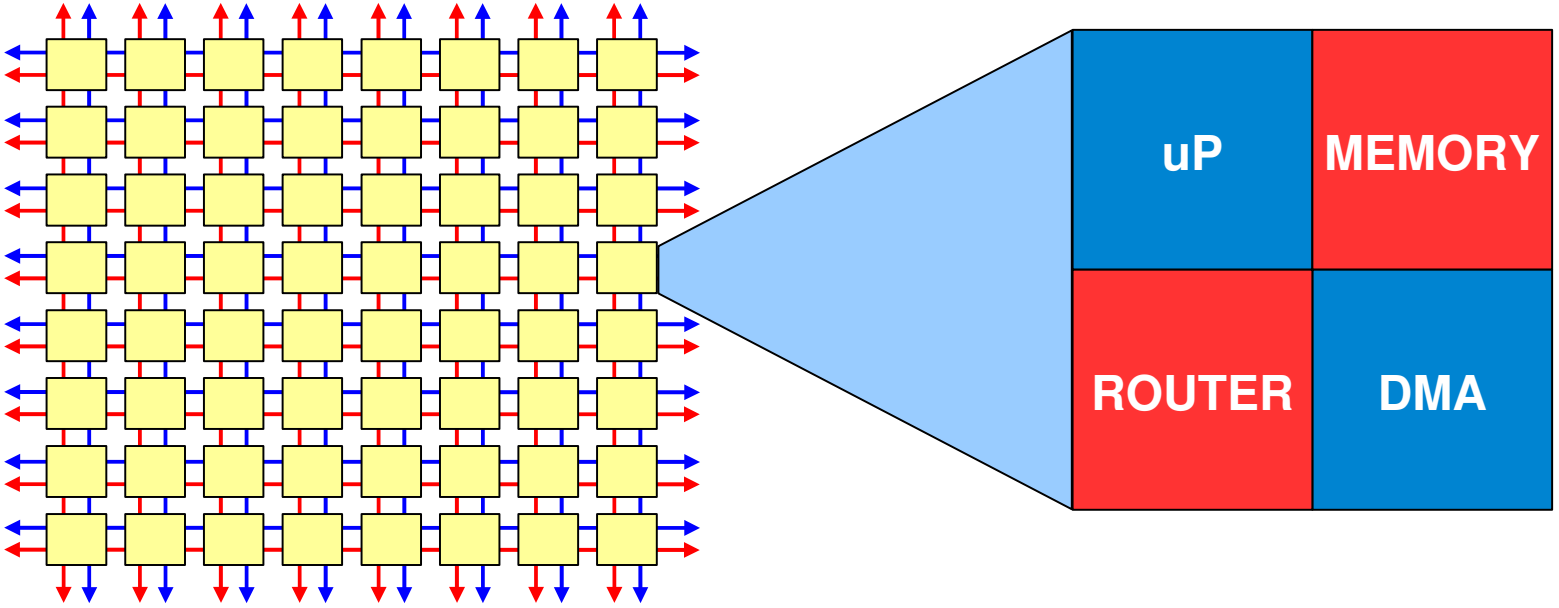
Microprocessor:
- Program Control
- Human interfaces,
- Knowledge extraction

SENSOR

FPGA

Peaceful coexistence between architectures...

Proposed Coprocessor Solution



- Mesh connected array of ANSI C-programmable processor cores
- 16 to1024 independent dual issue microprocessors

- Distributed Shared Memory Architecture
- Distributed DMAs
- IEEE 754 floating point support

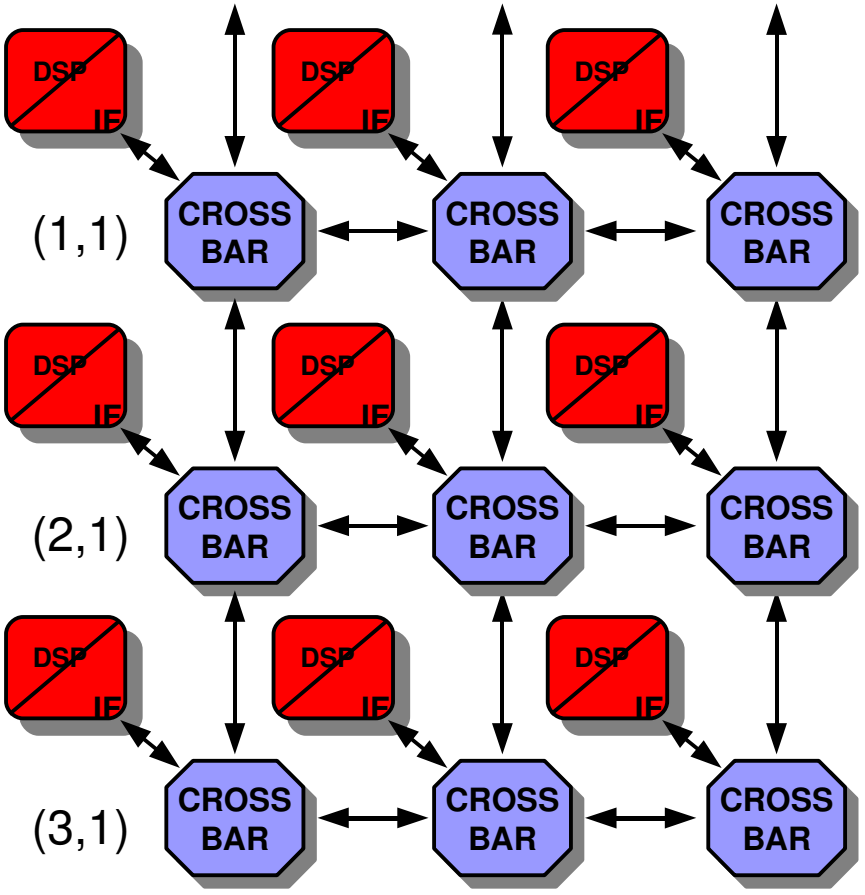
50 GFLOP/Watt Performance @ 65nm

The Processor

- Included:
  - ANSI-C programmability
  - Features that increased FLOPS/Watt
  - Floating point support
- Excluded:
  - Cache!!!
  - Compiler driven optimization
  - Instruction set optimization across a broad range of applications
  - SIMD

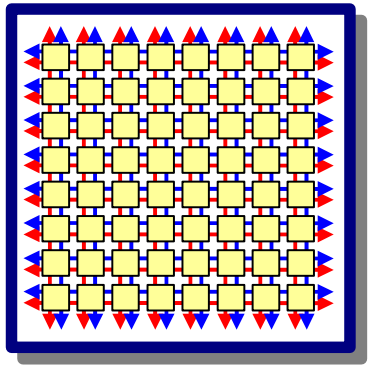
The Network

- Highlights:
  - On-chip wires are free!
  - Address used as 2D routing address
  - Bidirectional mesh network operating at same frequency as core
  - Optimized for deterministic data traffic and low latency communication
  - 64GB/sec BW at each node



Interfaces

- SOCs tend to have too many interfaces but never the right ones
- FPGAs can have the right interface and have hundreds of GPIO signals
- Use custom low power coprocessor-FPGA interface

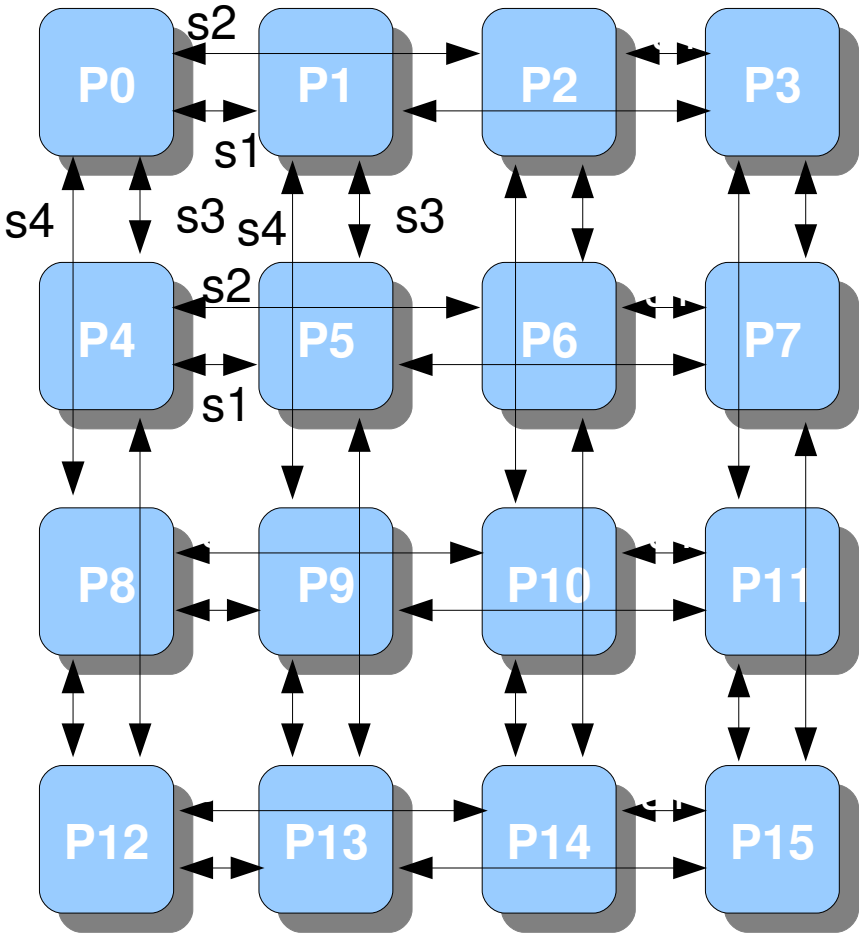


Custom
LVDS
Interface

FPGA

Multicore FFT Example

- **Approach:**
  - 1024 point FFT is spread over 16 processors
  - s1,s2,s3,s4 refer to the four FFT stages for combining data with 64 point complex data movements
  - Lower # procs transfer W0 to higher # procs.
  - Lower # proc calculates Wj0+Wj1 x Cj, higher # proc calculates Wj0-Wj1 x Cj
- **Results:**
  - NOC enables efficient multicore programming
  - < 3us execution time!
  - High efficiency
  - Work in progress, still room for improvement



Summary

- Heterogeneous computing is the only way to continue scaling performance
- ... but it will require some changes to the programming model
- 50 GFLOPS/Watt possible today in 65nm