The Raw Processor

A Scalable 32 bit Fabric for General Purpose and Embedded Computing

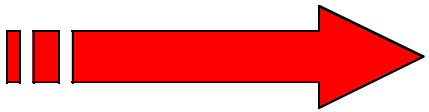
Michael Taylor, Jason Kim, Jason Miller,
Fae Ghodrat, Ben Greenwald, Paul Johnson,Walter Lee,
Albert Ma, Nathan Shnidman, Volker Strumpen, David Wentzlaff,
Matt Frank, Saman Amarasinghe, and Anant Agarwal

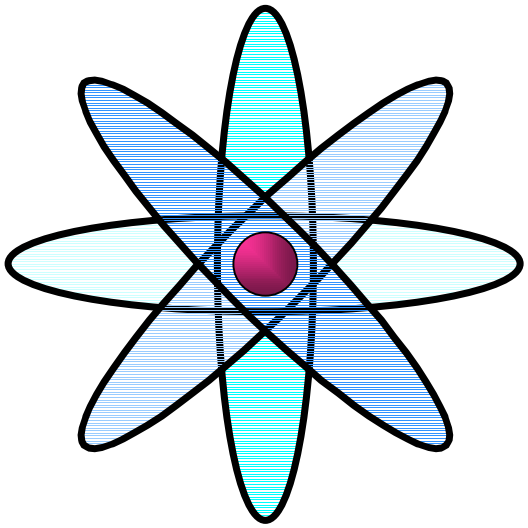
MIT Laboratory for Computer Science

http://cag.lcs.mit.edu/raw

Computer Architecture from 10,000 feet

```
foo(int x)
{  ..  }
```





class of computation

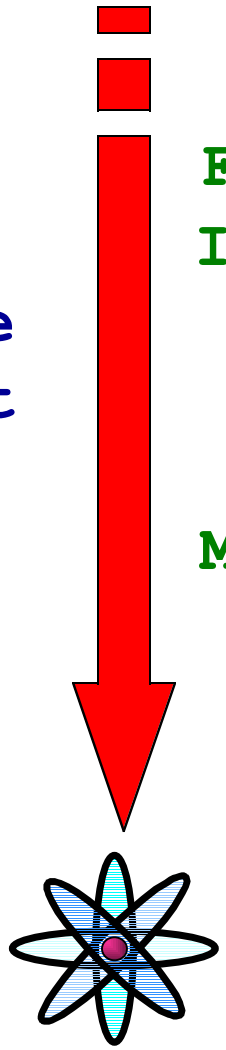
physical phenomenon

… we use abstractions to make this easier

The Abstraction Layers That Make This Easier

```
foo(int x) { .. }
```

Computation
Language / API
Compiler / OS
ISA
Micro Architecture
Floorplan / Layout
Design Style
Design Rules
Process
Materials Science
Physics



Fortran / Unix
IBM 360 /RISC/ Transmeta/x

Mead & Conway

Abstractions must change as world changes

Changing Applications

Language / API
Compiler / OS
ISA
Micro Architecture
Floorplan / Layout
Design Style
Design Rules
Process
Materials Science

Changes in physical constraints

More Wire Delay

More Gates Pins

More Power

Everyone is thinking about wire delay

Language / API	
Compiler / OS	
ISA	
Micro Architecture	**Partitioning(21264)**
	Pipelining (P4)
Floorplan / Layout	**Timing Driving Placement**
Design Rules	**Fatter wires**
Process	**Deeper wires**
Materials Science	**Cu wires**

The bottom line

Language / API
Compiler / OS
ISA

Micro Architecture

Floorplan / Layout
Design Rules
Process
Materials Science

Raw tries to solve *these* problems by exposing the underlying resources with a scaleable, parallel ISA.

It orchestrates these resources with spatially-aware compilers.

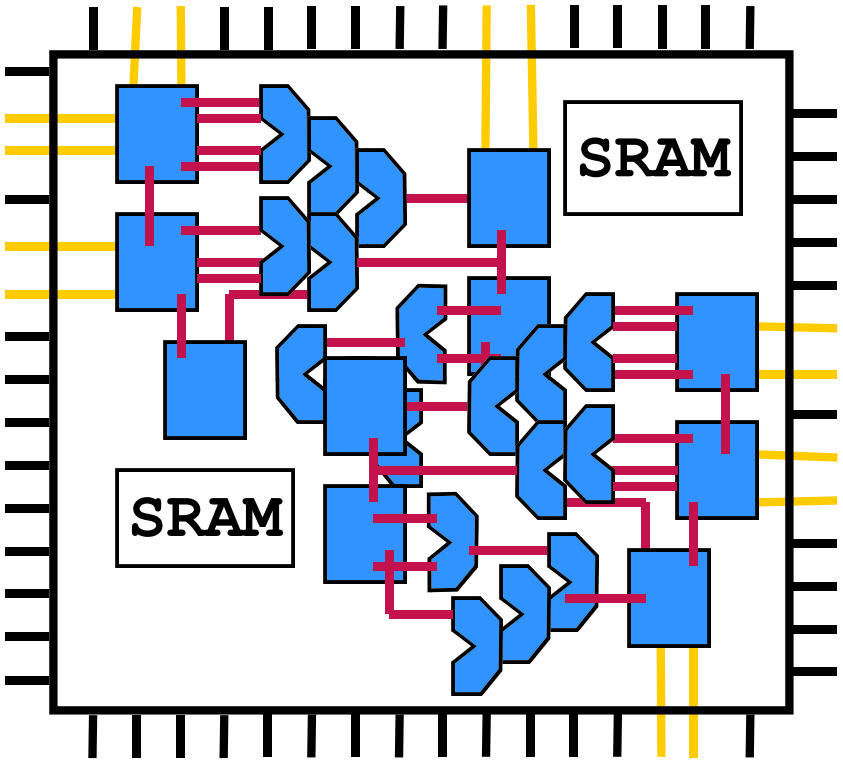
More	More	More
Wire	Gates	Power
Delay	Pins	

Intuition

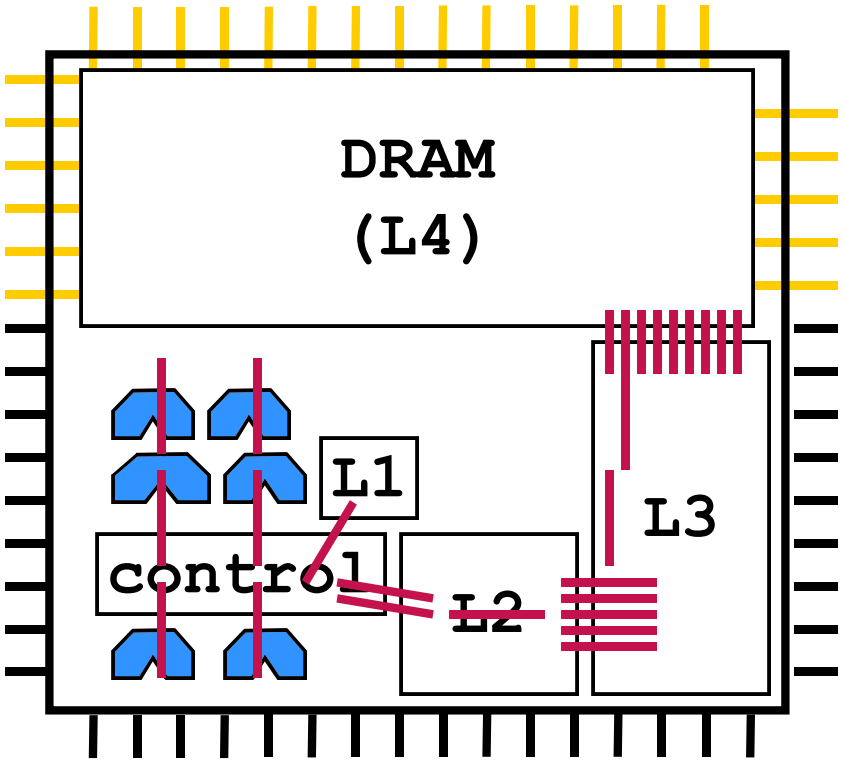
1 cycle wire

Customized gates
Customized wiring
Customized placement
Customized pins
Fixed function

Heroic attempts to distribute computation across a hand-full of relatively nearby ALUs.
I/O through memory interface



Custom Hardware



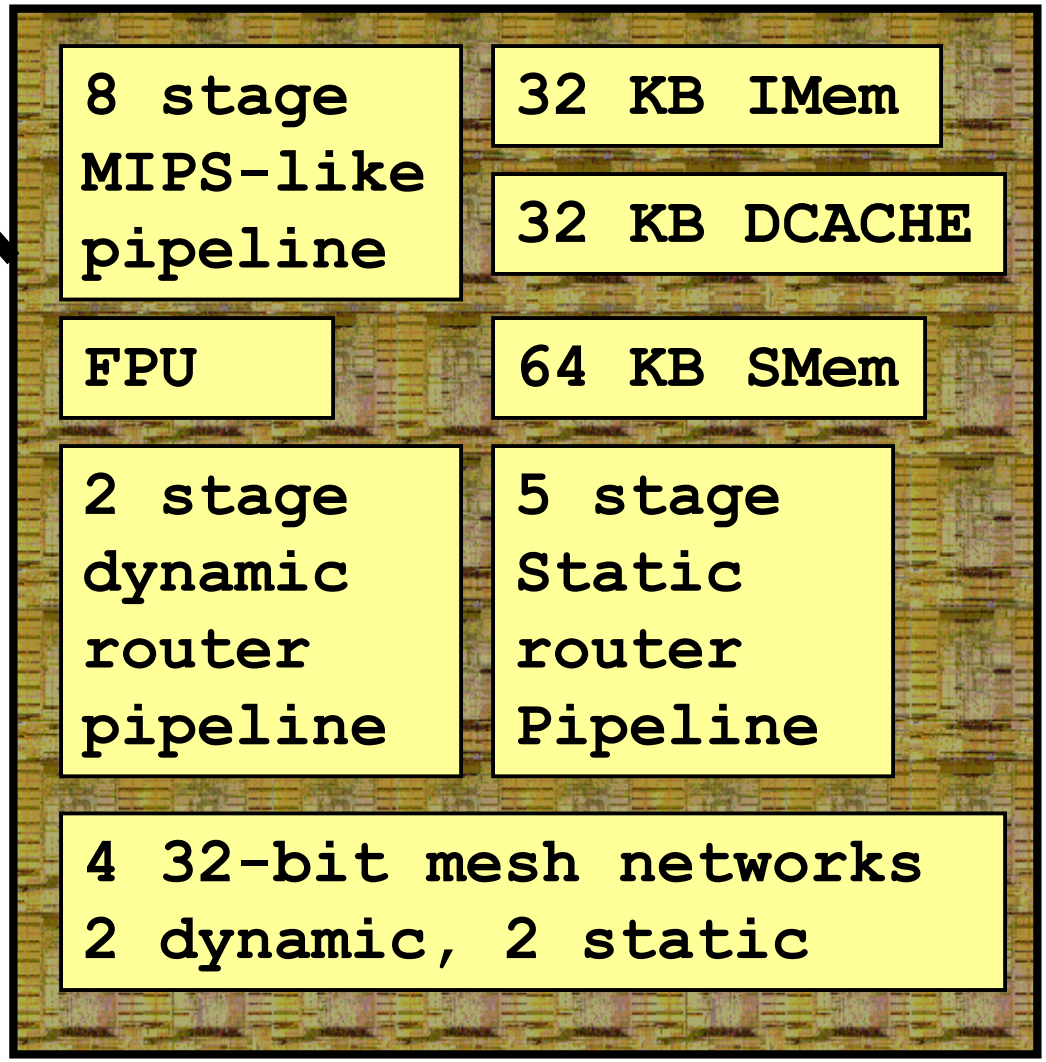
Monolithic ISA

The Raw ISA provides a parallel interface to the gates

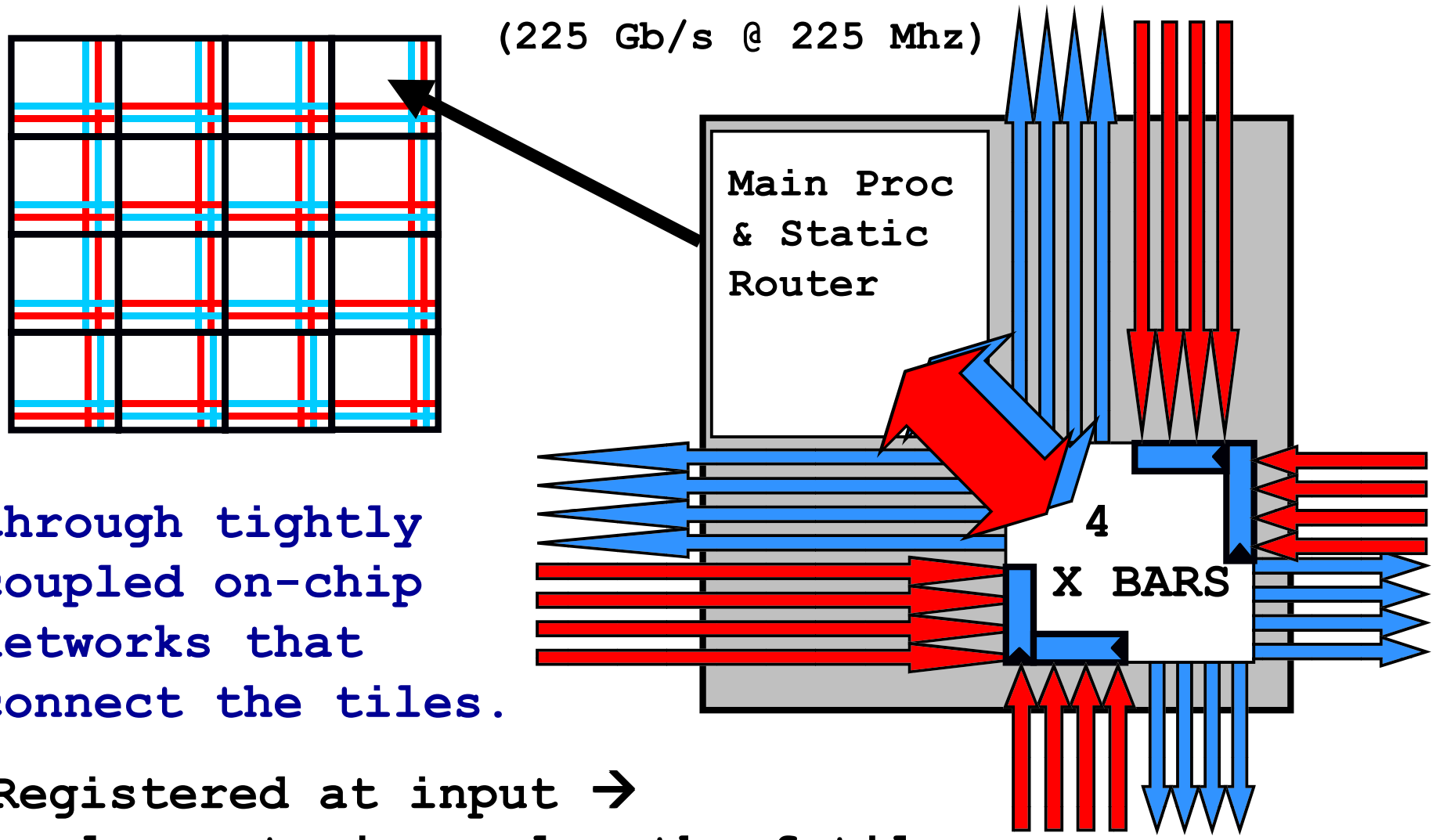
Raw is an array of replicated tiles.

Use more tiles to get more computation.

Tile processor



The Raw ISA provides a parallel interface to the wiring resources



through tightly coupled on-chip networks that connect the tiles.

Registered at input →
longest wire = length of tile

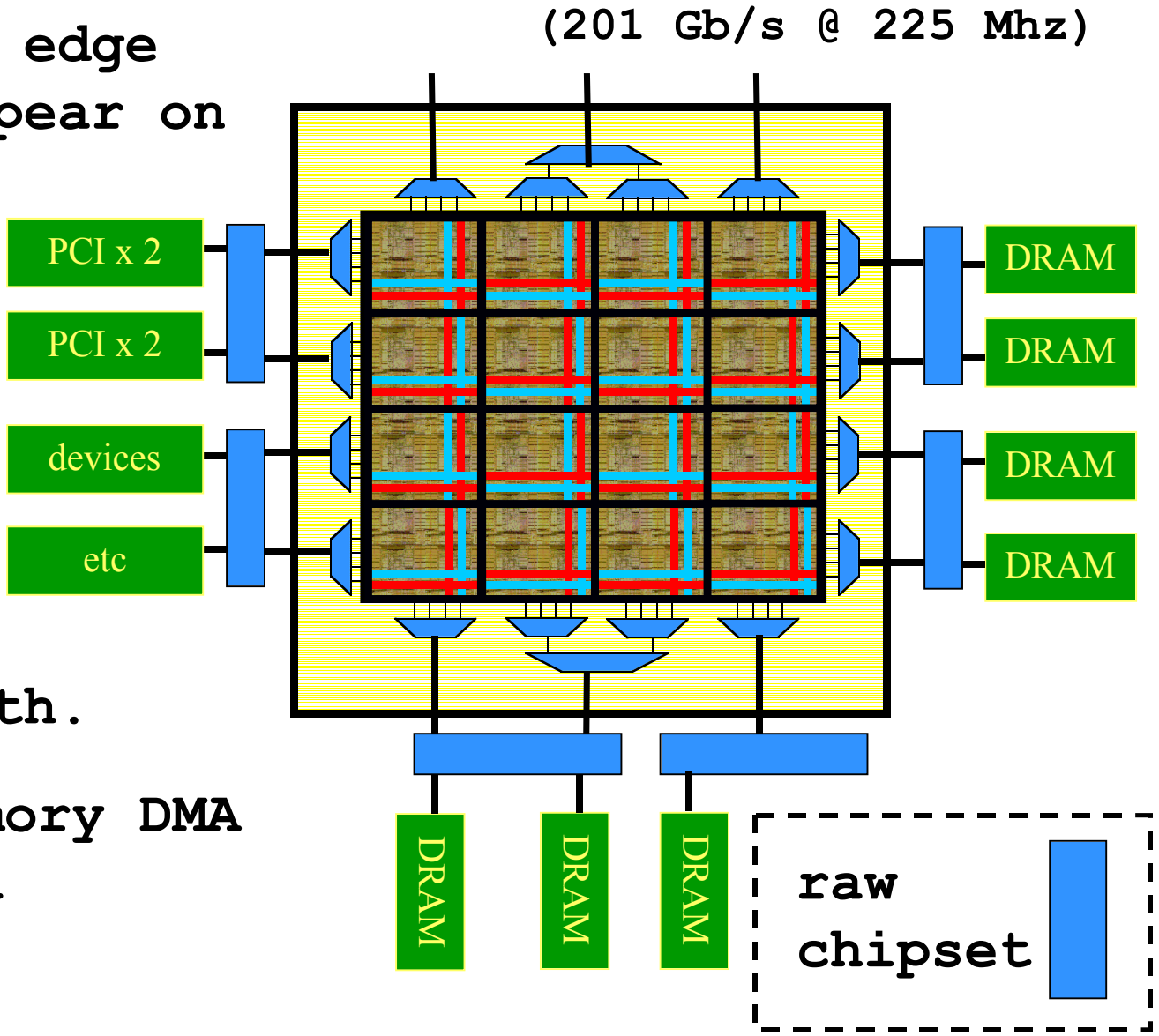
The Raw ISA provides a parallel interface to the pins

Routes off the edge of the chip appear on the pins.

Pins run at chip speed.

Gives user direct access to pin bandwidth.

Device <-> Memory DMA routed through the raw chip.

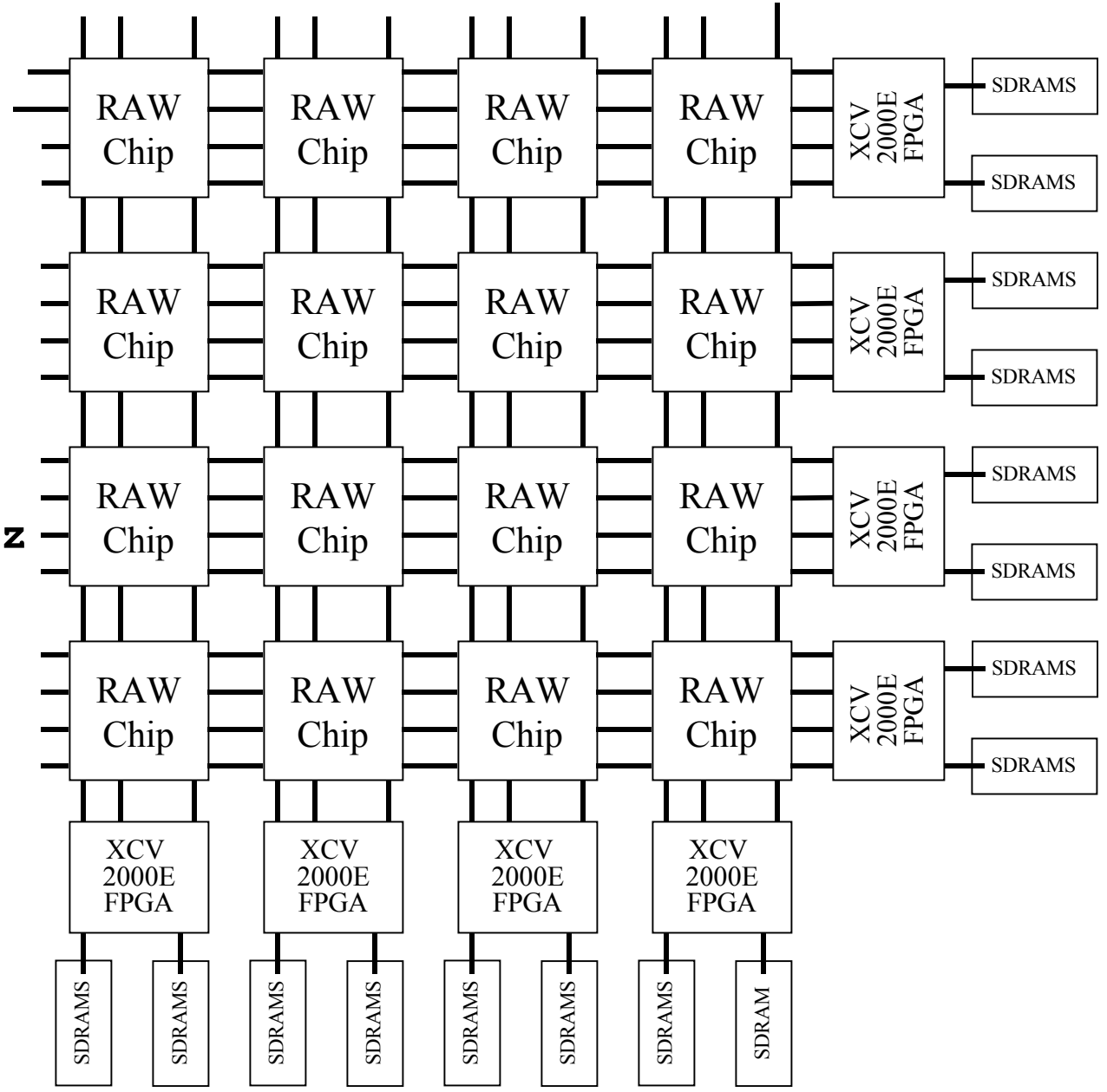


The Raw ISA is scalable

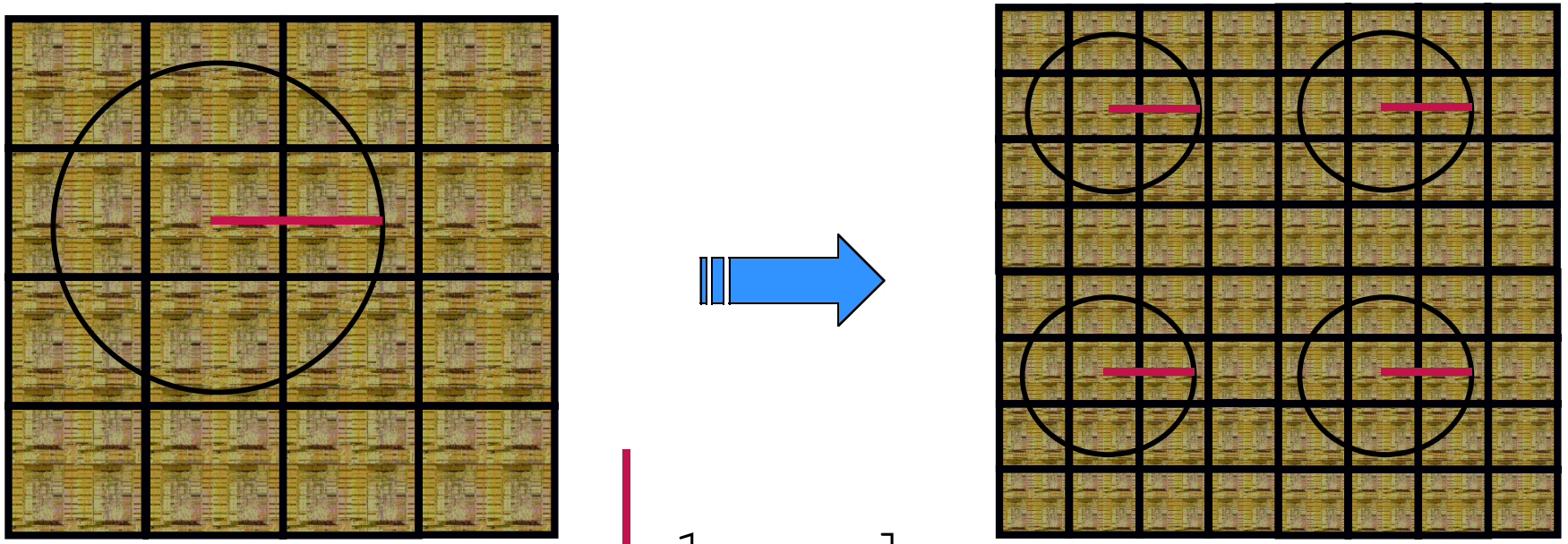
Simulates larger Raw processors (to 64 chips) at full speed.

1 16 Chip board:

256 Tiles
57 GFLOPS @ 225 MHz
32 MB SRAM onchip
806 Gb/s I/O
32 KB of RF



QED: The Raw ISA scales



.15u

.07u

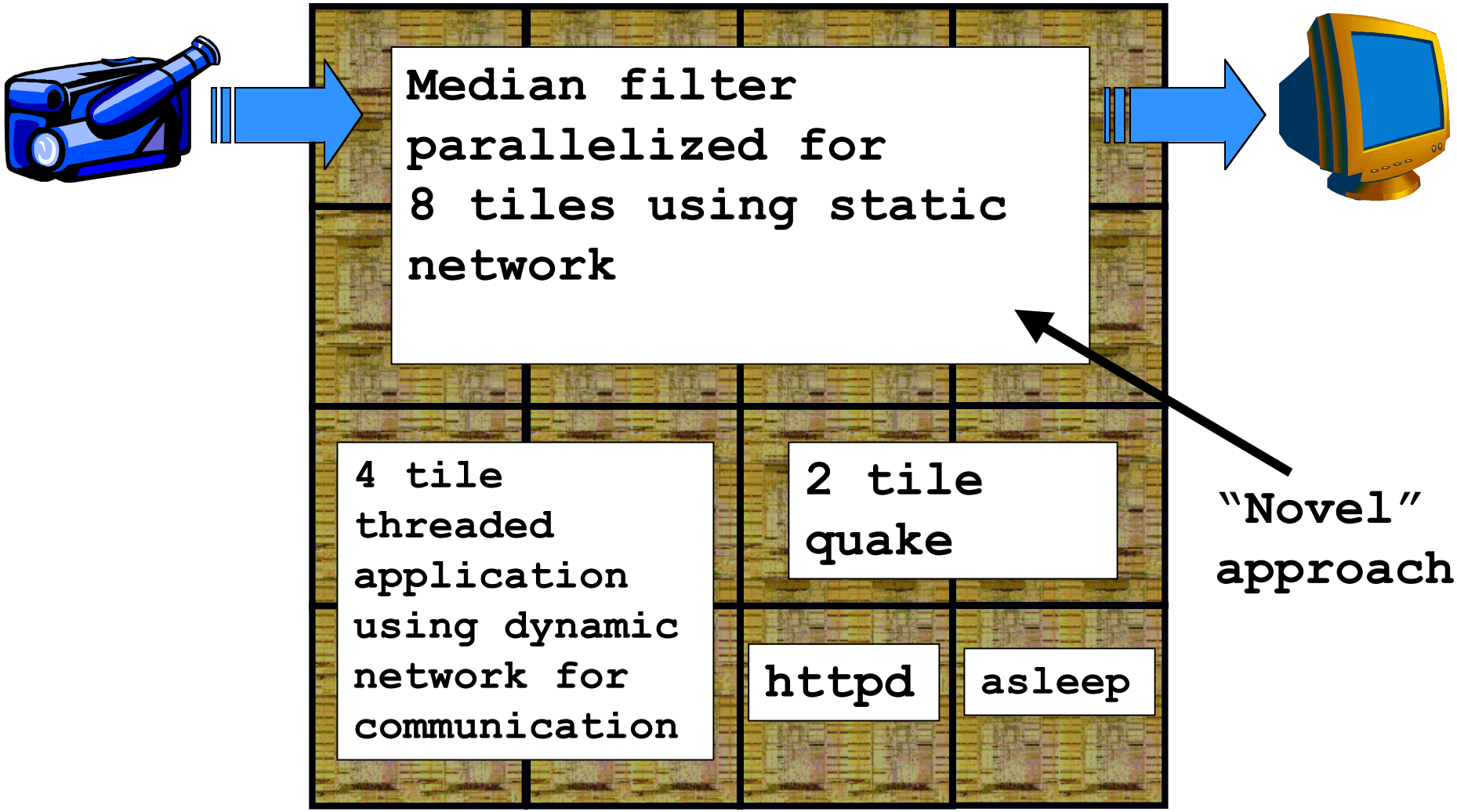
 # tiles and I/O ports scale linearly with gates and pins.

1. **Wire delay**
2. **Design complexity**
3. **Verification complexity**

… are all independent of transistor count.

Raw is also backwards-compatible.

Raw: How tiles are used

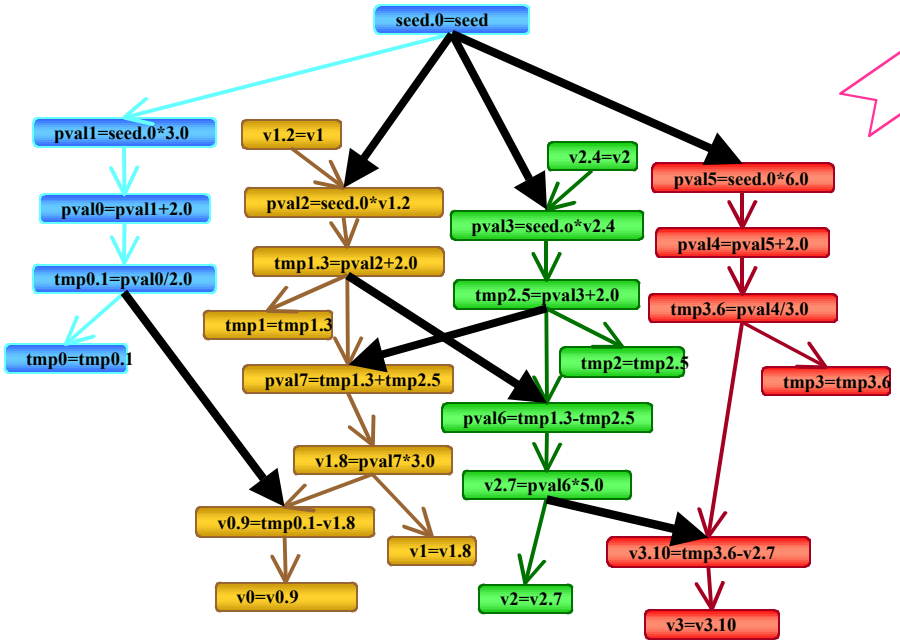


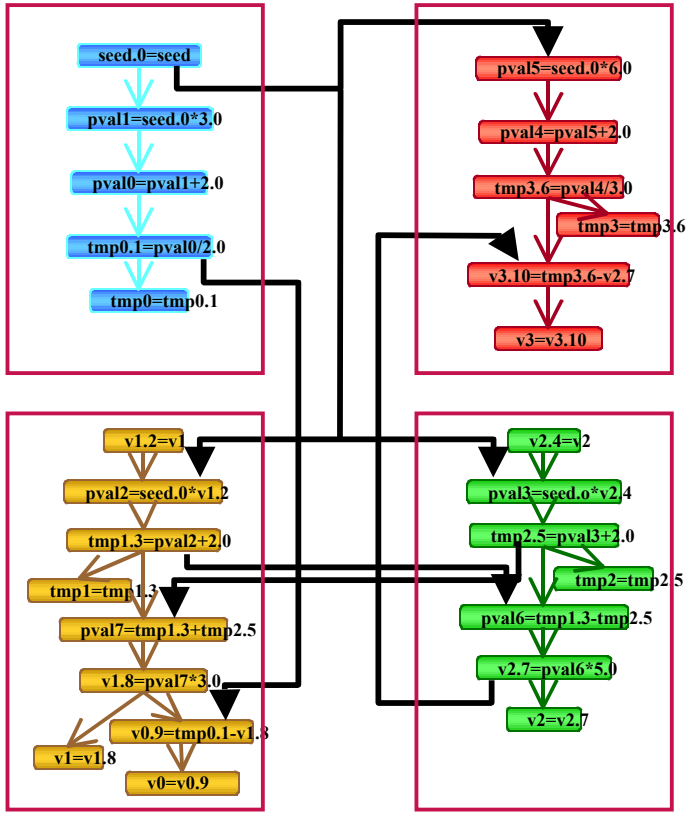
```
tmp0 = (seed*3+2)/2
tmp1 = seed*v1+2
tmp2 = seed*v2 + 2
tmp3 = (seed*6+2)/3
v2 = (tmp1 - tmp3)*5
v1 = (tmp1 + tmp2)*3
v0 = tmp0 - v1
v3 = tmp3 - v2
```

RawCC Operation:

Parallelizes C code onto static network

Black arrows = Static Network



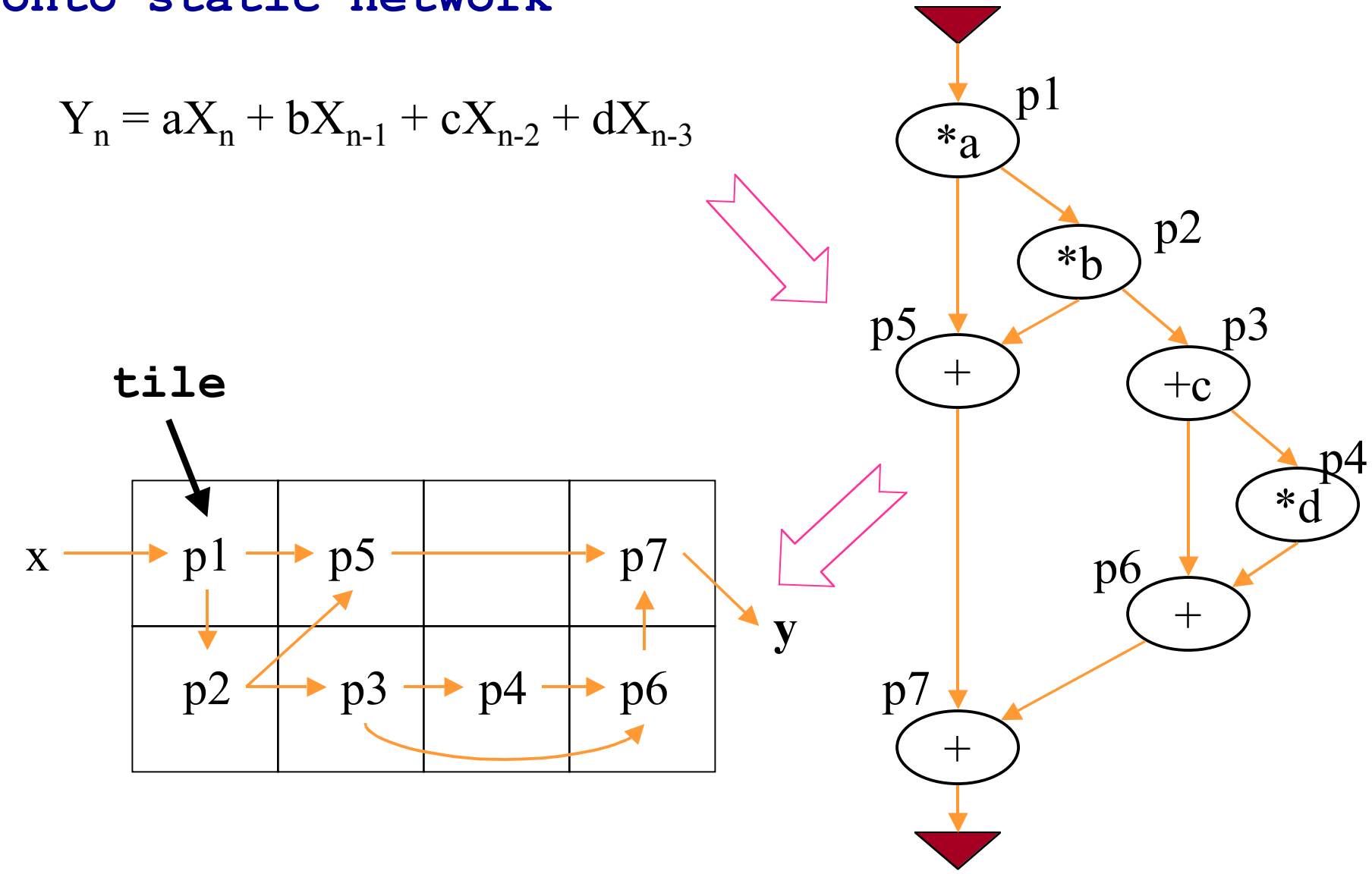


Low Network latency important.

Raw Stream Compiler

Maps pipeline parallel code onto static network

$$Y_n = aX_n + bX_{n-1} + cX_{n-2} + dX_{n-3}$$



Enabler: The Raw Networks

The Raw ISA treats the networks as first class citizens, just like registers.

ALU-ALU latency in cycles:

Dynamic Networks: 5 + # hops + # turns

1. dimension-ordered, worm-hole routed
2. for cache misses / IO / interrupts and unpredictable communication

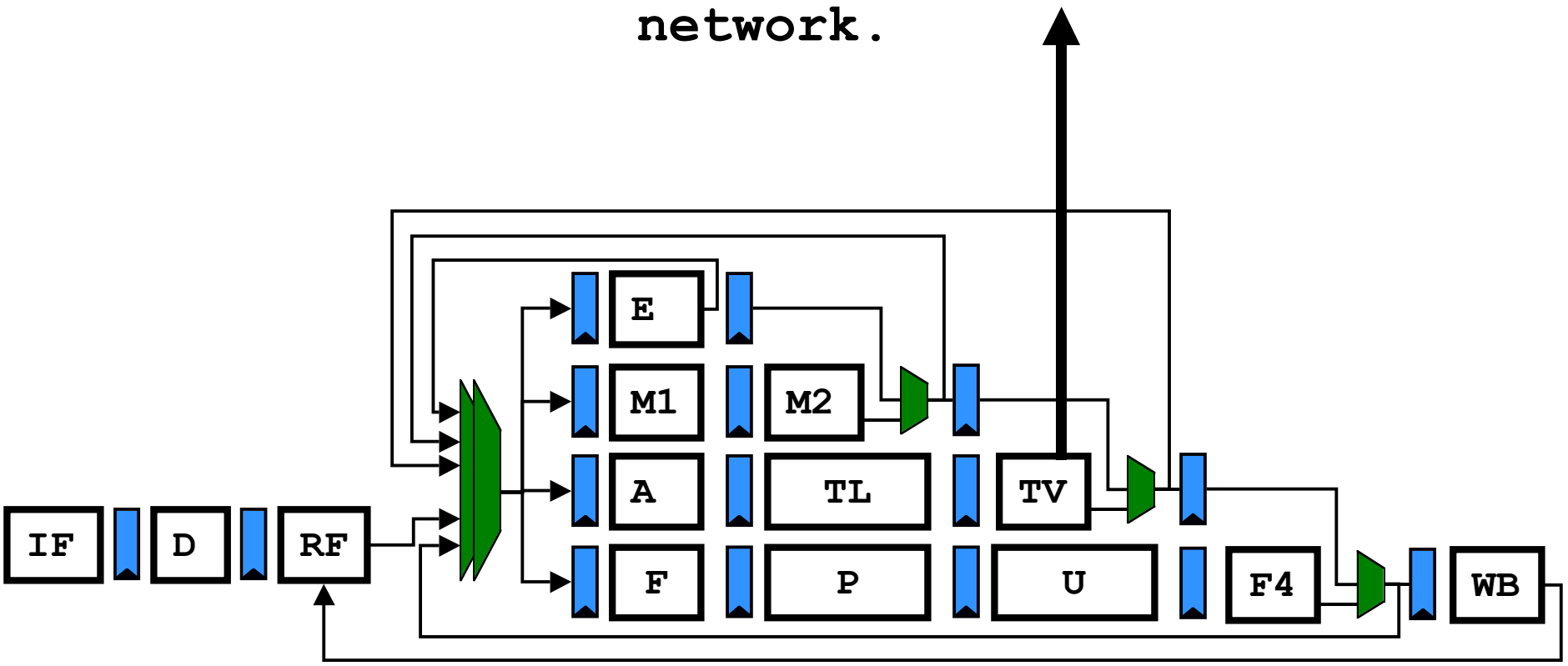
Static Networks: 2 + # hops

1. routes compiled into static router SMEM
2. Messages arrive in known order

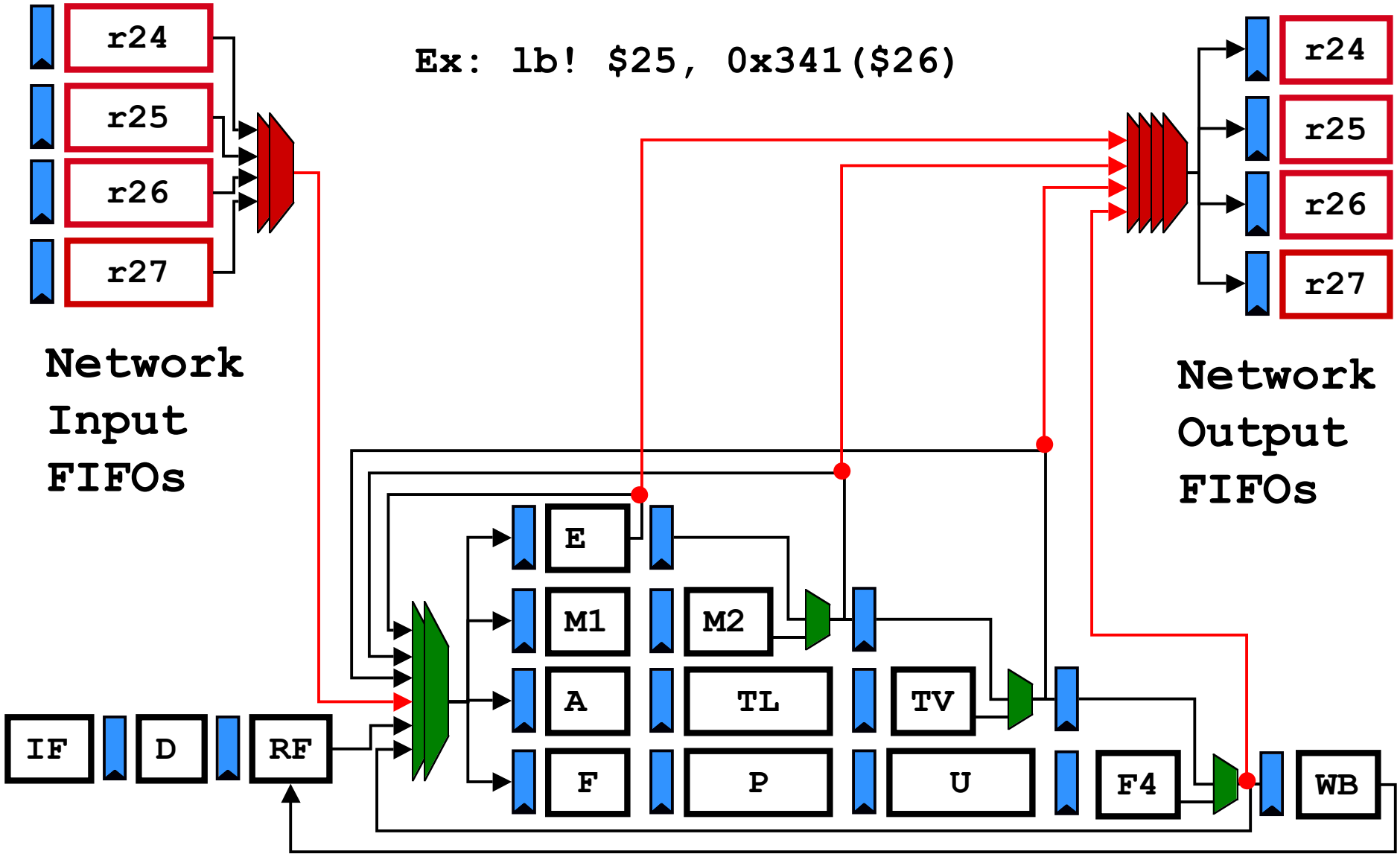
Throughput: 1 word/cycle per dir. per network

This network is not a first class citizen.

To other tiles, through memory system that happens to go over a network.



This is: the networks are tightly coupled into the bypass paths



Raw Stats

IBM SA-27E .15u 6L Cu

18.2mm x 18.2mm die.

.122 Billion Transistors

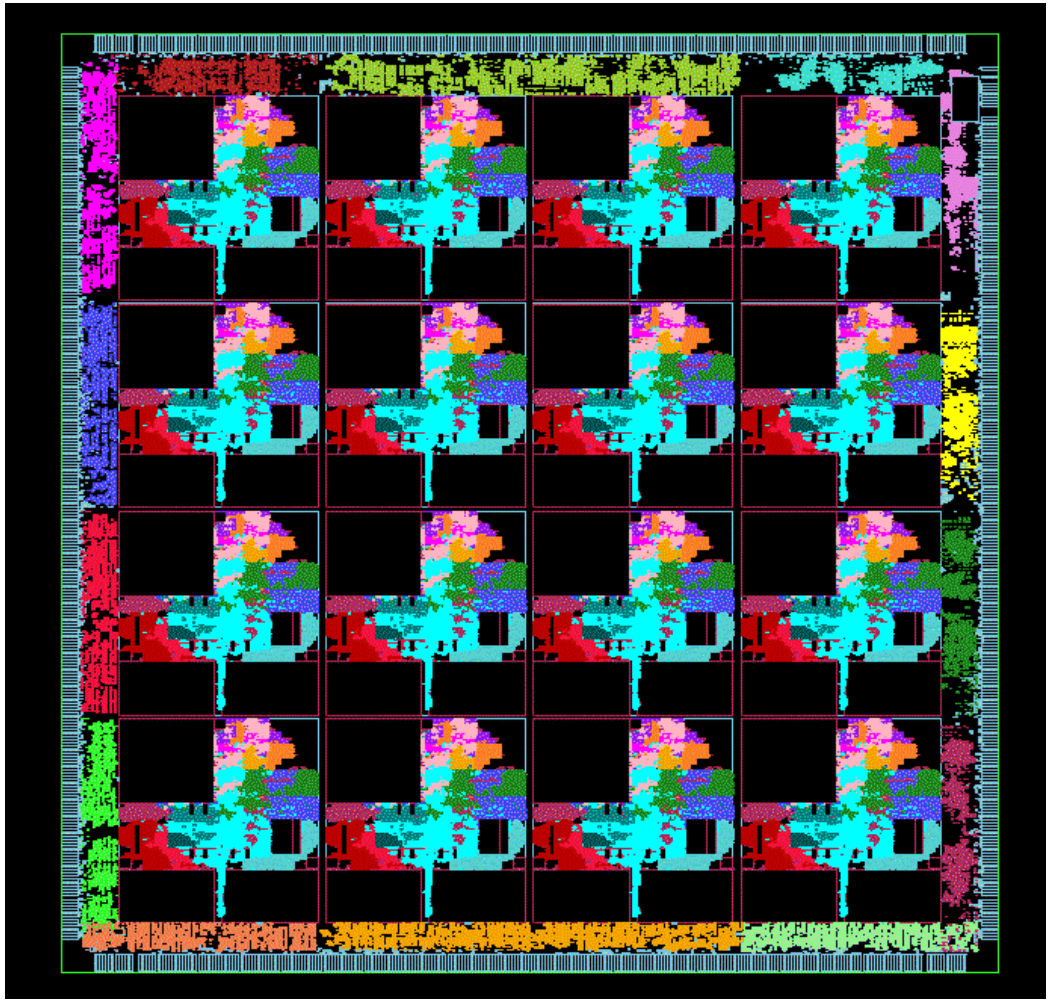
16 Tiles

2048 KB SRAM Onchip

1657 Pin CCGA Package
(1152 HSTL signal IO)

~225 MHz

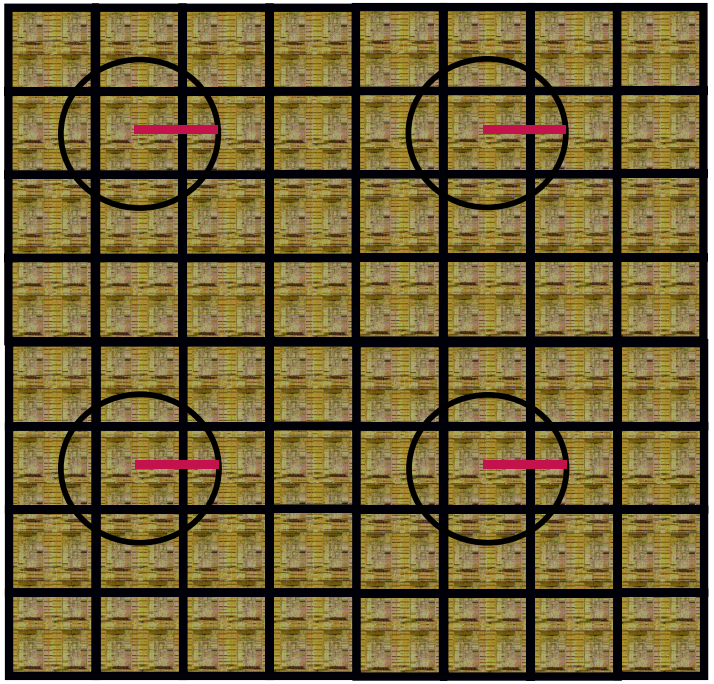
~25 Watts



For architectural details, see:
http://cag.lcs.mit.edu/pub/raw/documents/RawSpec99.pdf

Summary

Raw exposes wire delay at the ISA level. This allows the compiler to explicitly manage gates in a scaleable fashion.



Raw provides a direct, parallel interface to all of the chip resources: gates, pins, and wires.

Raw enables the use of these gates by providing tightly coupled network communication mechanisms in the ISA.