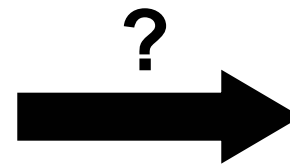
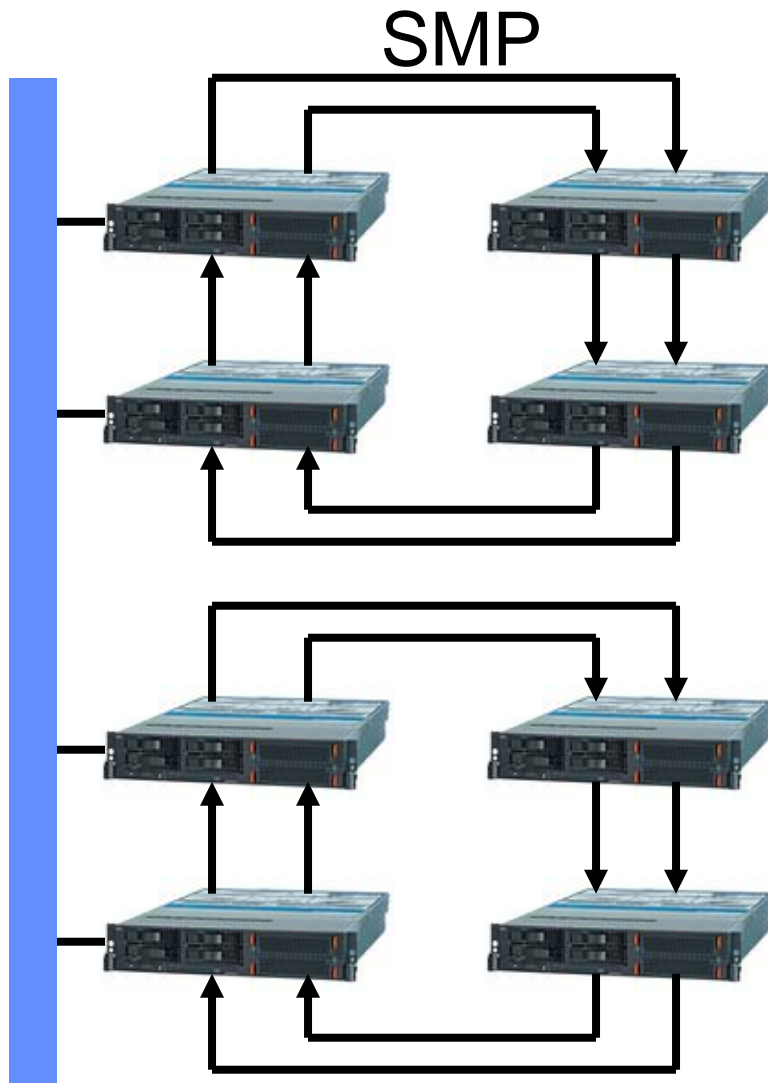

Design and Integration Challenges for On-Chip Networks

Steve Keckler

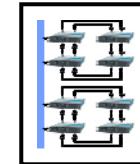
Department of Computer Sciences
The University of Texas at Austin



From System to On-chip Networks



CMP

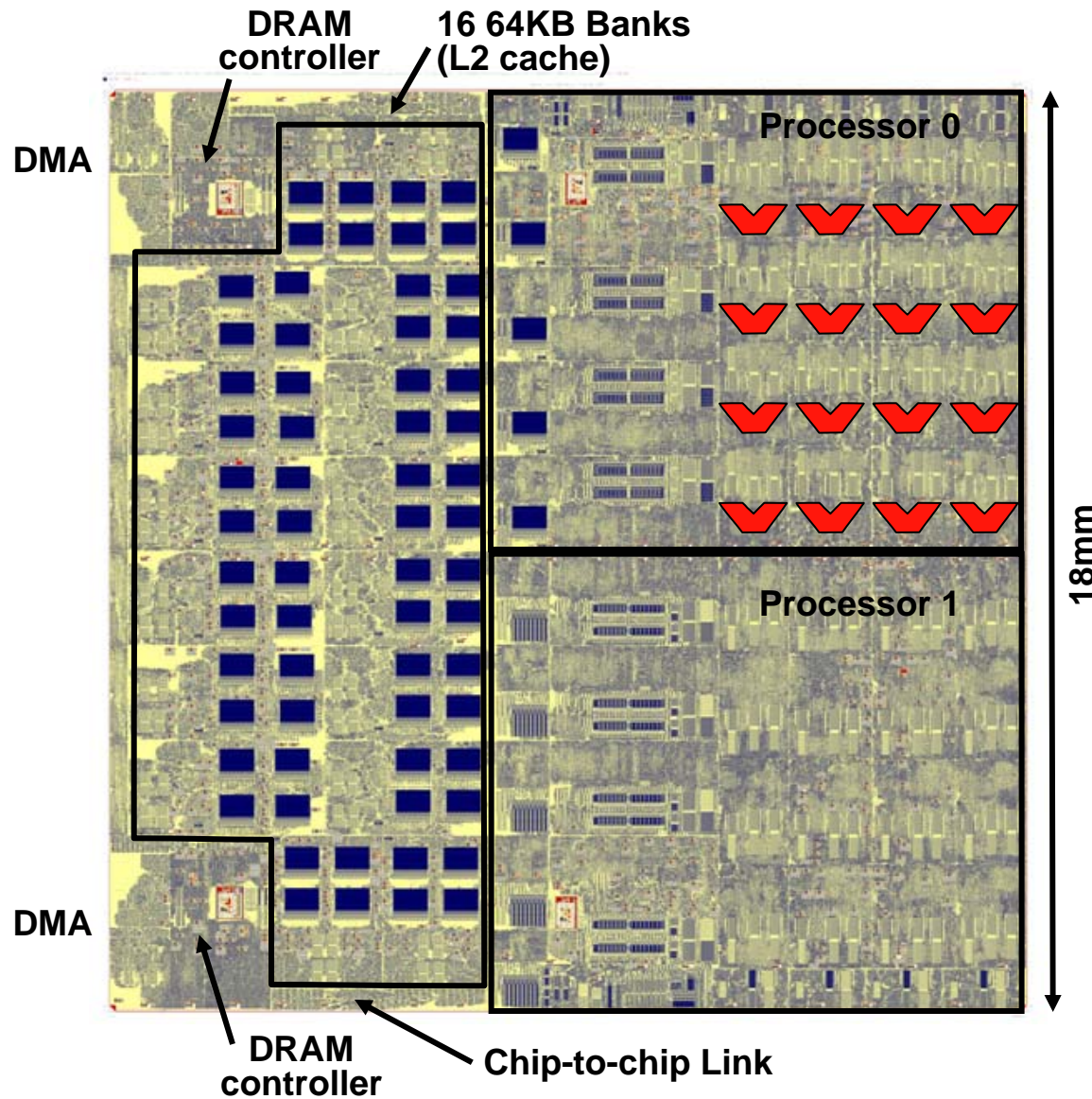


- What's different?
 - More wires
 - May run at high (processor) clock rate
 - Less power per transmitted bit
 - Topology restricted
- More opportunities and challenges

Opportunities for On-chip Networks

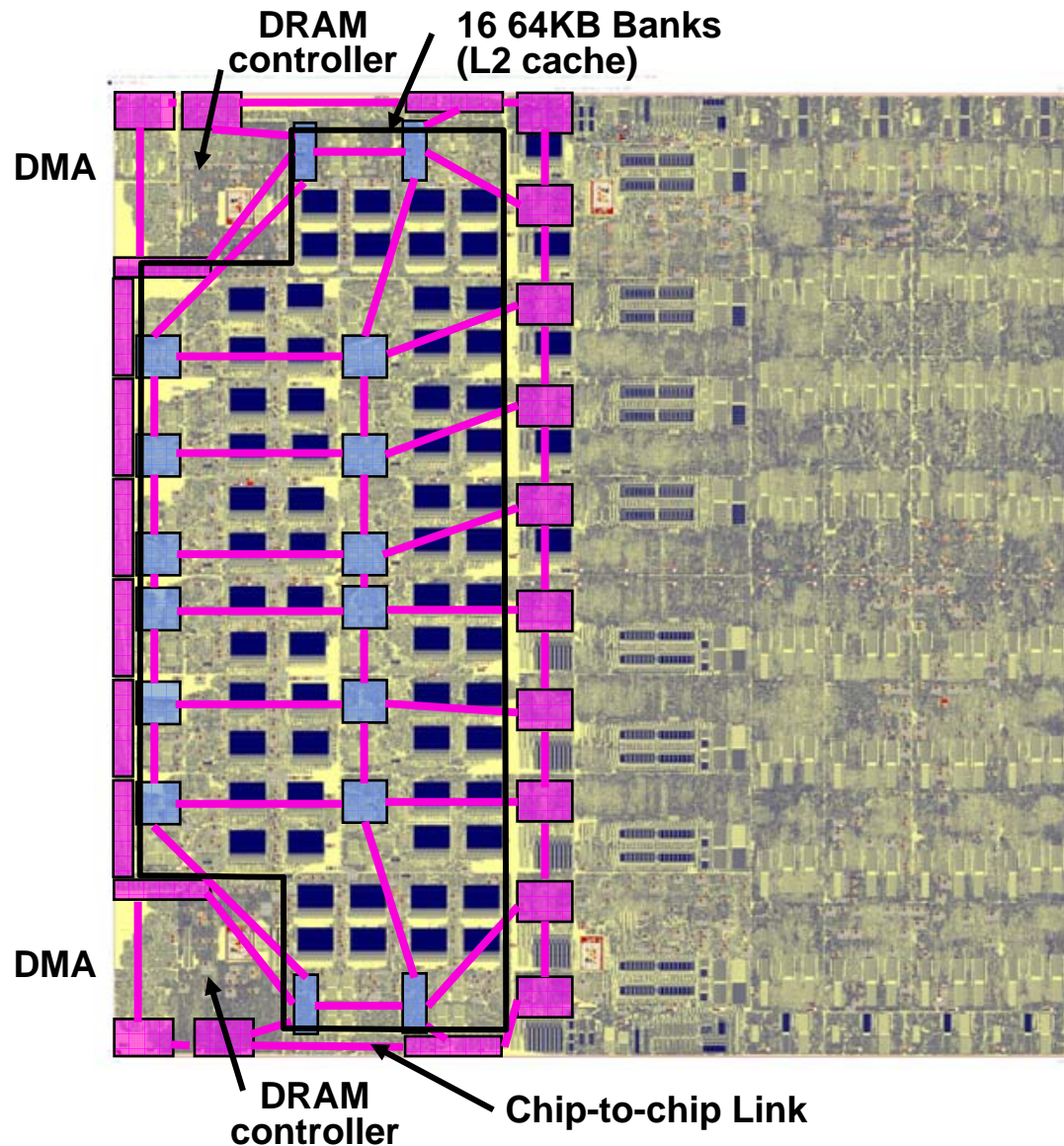
- Integration
 - Can tightly couple with computation/memory elements
 - High bandwidth easy with wide links
 - Control networks, operand networks, memory networks
 - Better than busses, can exploit physical locality
- Extensibility
 - Easy to integrate different types of components
- Configurability - change behavior of hardware
 - Tolerate hard faults
 - Adapt system to application needs
- Critical considerations
 - Latency
 - Minimize time in routers
 - Lightly loaded latency can be important
 - Area/power must be budgeted
 - mm² and watts to interconnect comes from other on-die components

TRIPS Tiled and Networked Processor



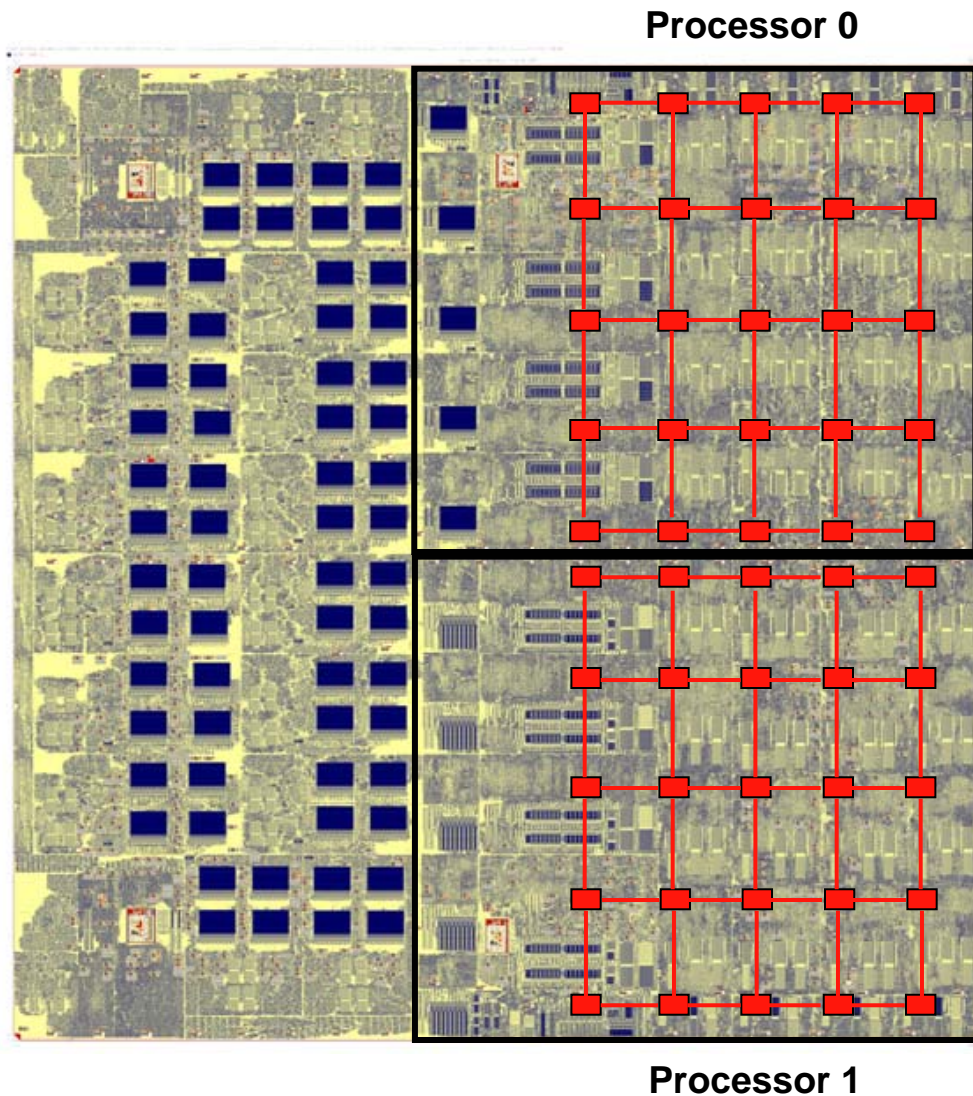
- SOC-like design style
 - Individually designed tiles
 - 6-8 mm² each
 - 170M transistors
- Networks
 - Memory
 - Operands
 - Control
- Networks enable
 - Distributed and scalable design
 - Fast design cycle

TRIPS Memory Network (OCN)



- 4x10 mesh
 - router embedded in memory tile
 - network interface (router + routing table)
 - 128-bit channels/flits
 - 4 VCs for 4 priorities
 - 1 cycle per hop
 - Runs at processor clock
 - Bisection BW 64 GB/sec at 500MHz
- Connects heterogeneous memories and controllers to each other and procs
 - Near banks are faster to access
- Configurable translation
 - Cache organization
 - Cache vs. local memory

TRIPS Operand Network (OPN)



- 5x5 mesh network
 - 140 bit channels
 - No VCs
 - Bisection BW 80GB/sec at 500MHz
 - 1 cycle per hop
 - Optimistic header injection
- Connects ALUs to each other and to L1 cache
- Tightly integrated into processor core
 - Takes place of bypass bus

Design Observations

- Networked “SOC” design style really worked
 - Large chip designed and verified quickly by small team
 - Configurability of OCN was simple
- Latency matters...a lot
 - Processor performance drops by 30% at 2 OPN hops per cycle
 - Traffic alternates between low and high load
 - Cannot sacrifice low-load latency for high-load throughput
- Timing - control path latency is 2x datapath
 - Opportunities for custom circuits
- Area adds up quickly
 - OCN + OPN = 20% of chip area
 - Reducing OCN flit buffers to minimum size was fine
 - Keep designs simple for speed and area

Future Challenges

- Scaling to hundreds/thousands of on-chip cores/tiles
 - Yet maintain low latency
 - Need alternatives to move control further off critical path
 - But without adding too much complexity
 - Clocking - synchronous vs. asynchronous
 - Asynchronous attractive, but may not be able to afford resynchronization costs
 - Topology - low vs. high dimension networks
 - Must maintain low latency to near neighbors
 - Design efficiency - power and area
- Reliability - not too worried, keep off critical path
- Benchmarking - what is a realistic workload?
- Extensibility - how to integrate w/ off-chip network
- Standards - would enable better design modularity
 - Connect SOC components from different designers (Arteris)
 - Classic tradeoff between generality and performance

**“It’s tough to make predictions,
especially about the future.”**

- Yogi Berra