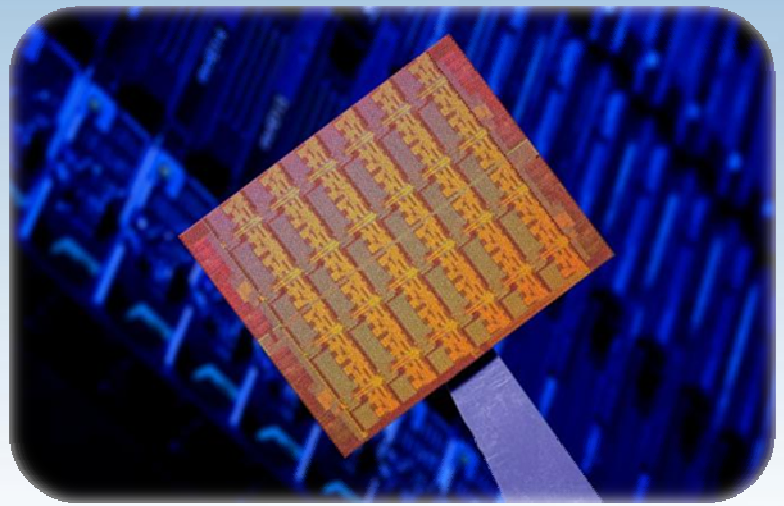
“A 48-Core Single Cloud Computer”

Intel’s experimental many-core processor



Jason Howard
Advanced Microprocessor Research
Intel Labs

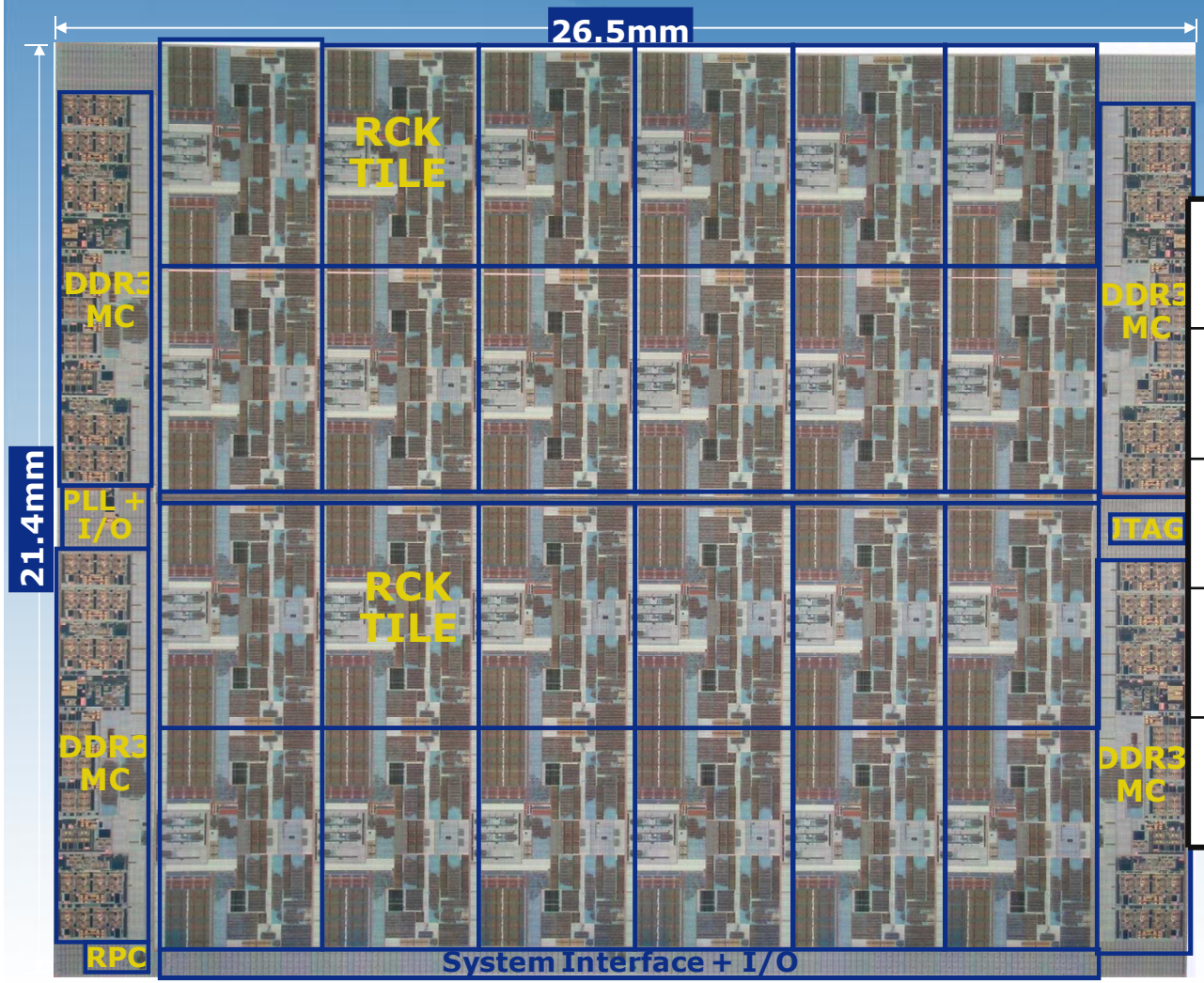
Content

- Feature set
- Architecture overview
  - Core
  - Memory Model
  - Interconnect Fabric
  - I/O and System Overview
- Design Overview
  - Clocking
  - Message Passing
  - Power management
- Silicon Results
- Programming
- Summary

SCC Feature set

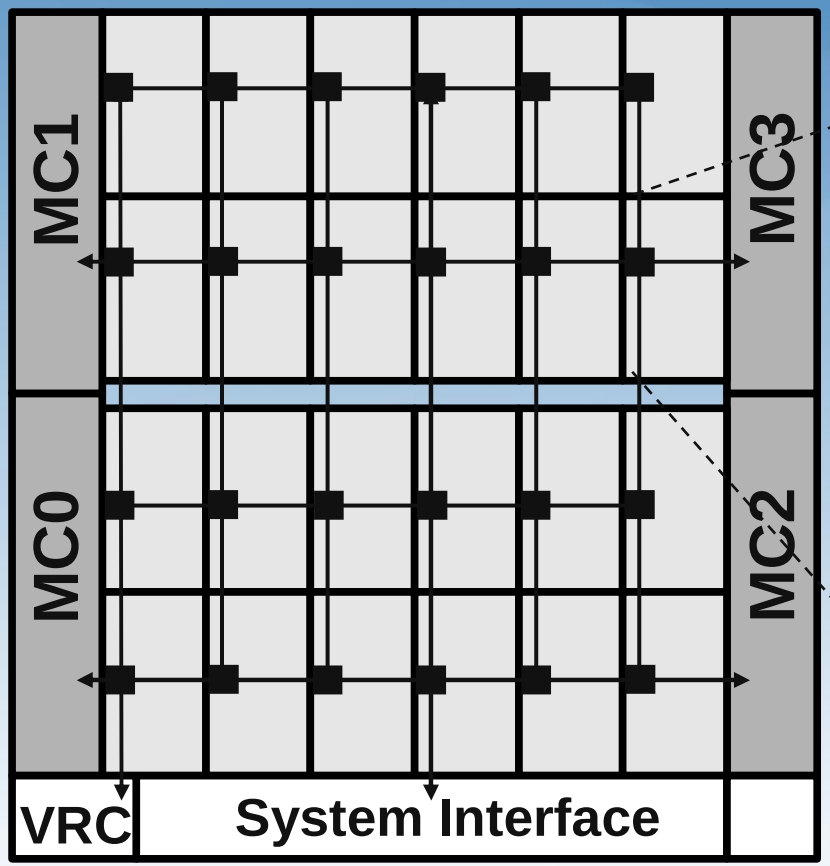
- First Si with 48 iA cores on a single die
- Next generation 2D mesh interconnect
  - Bisection B/W 1.5Tb/s to 2Tb/s, avg. power 6W to 12W
- Power envelope 125W
  - Core @ 1GHz, Mesh @ 2GHz
- Message passing architecture
  - No coherent shared memory
  - Proof of Concept for scalable solution for many core
- Fine grain dynamic power management
  - On die controller for on-package VRs
  - Frequency modulation

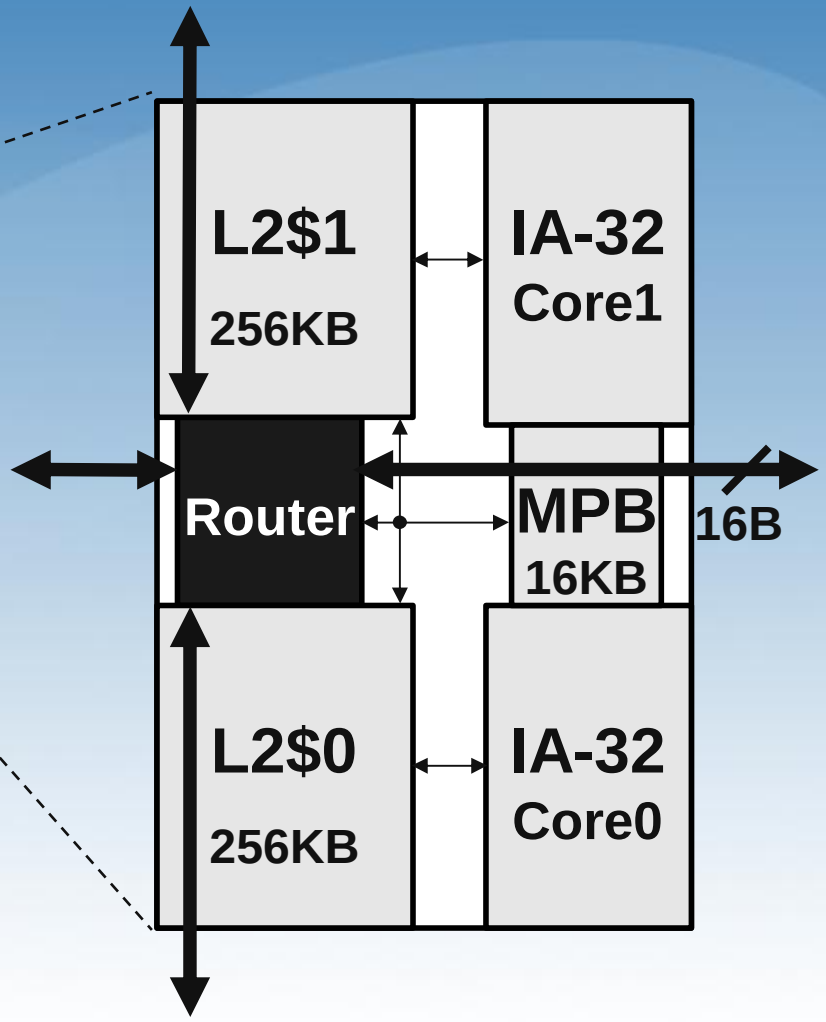
SCC Fullchip



Technology	45nm Hi-K CMOS Process
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	Die: 1.3B, Tile: 48M
Tile Area	18.7mm^2
Die Area	567.1mm^2

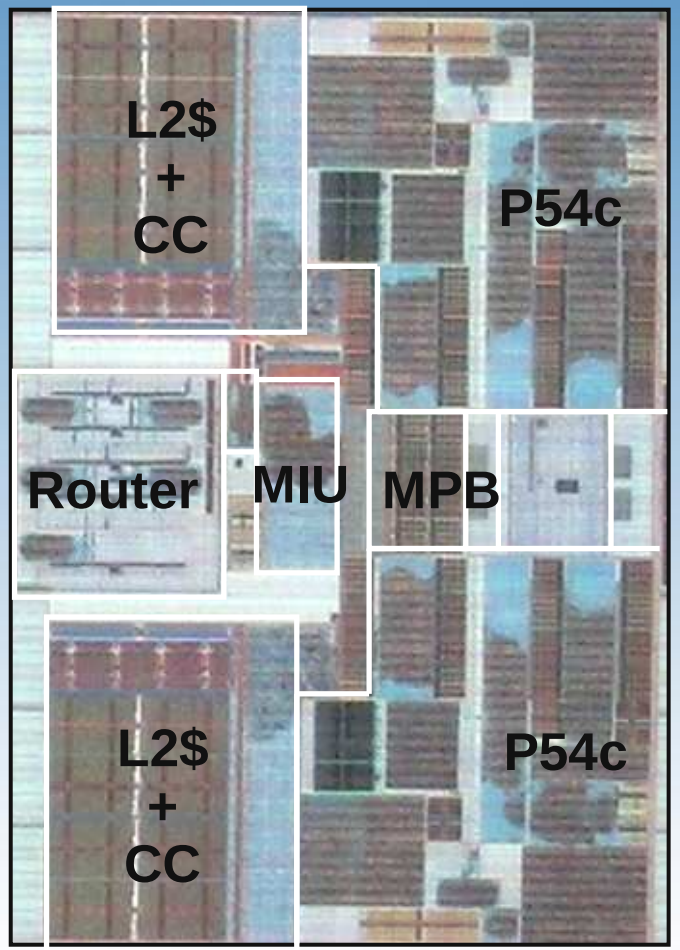
Die Architecture





2 core clusters in 6x4 2-D mesh

SCC Tile



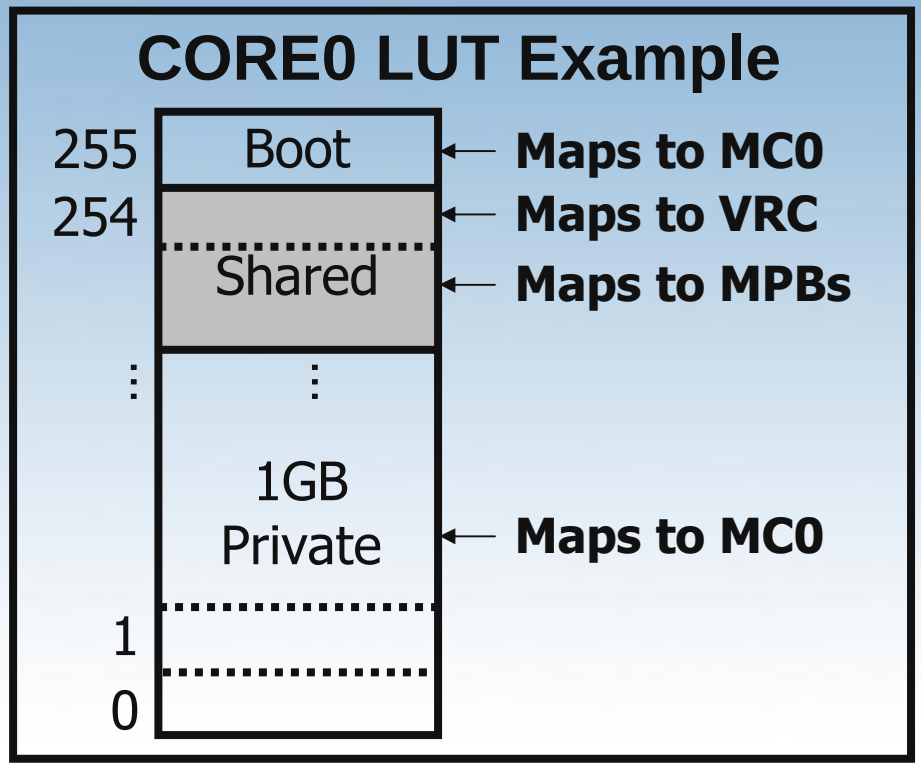
- 2 P54C cores (16K L1$/core)
- 256K L2$ per core
- 16K Message passing buffer
- Mesh Interface Unit
  - Imbedded configuration registers
- Router

- Tile area 18.7mm2
- Core area 3.9mm2

intel

Core Memory Management

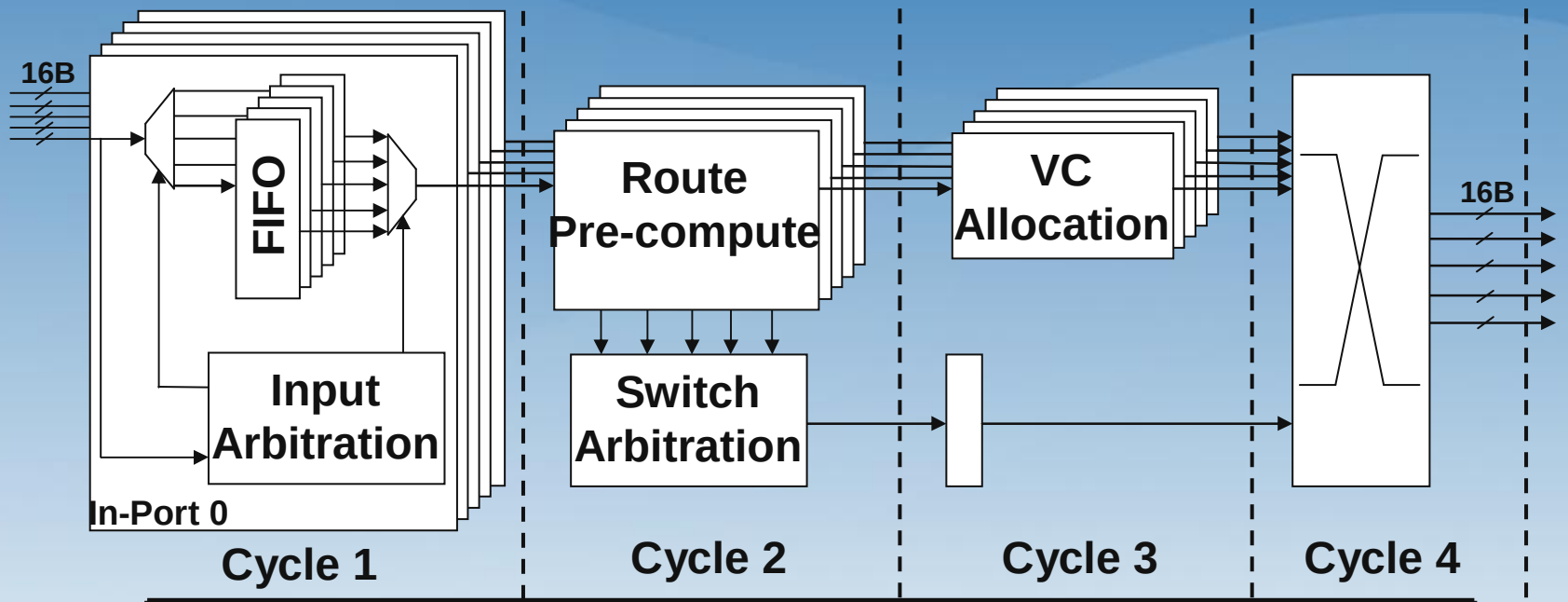
- Each cores is allocated independent, private memory
- Core cache coherency is restricted to private memory space
  - Maintaining cache coherency for shared memory space is under software control
- Each core has an address Look Up Table (LUT) extension
  - Provides address translation and routing information
- LUT values must fit within the core and memory controller constraints
- LUT boundaries are dynamically programmable



On-Die 2D Mesh

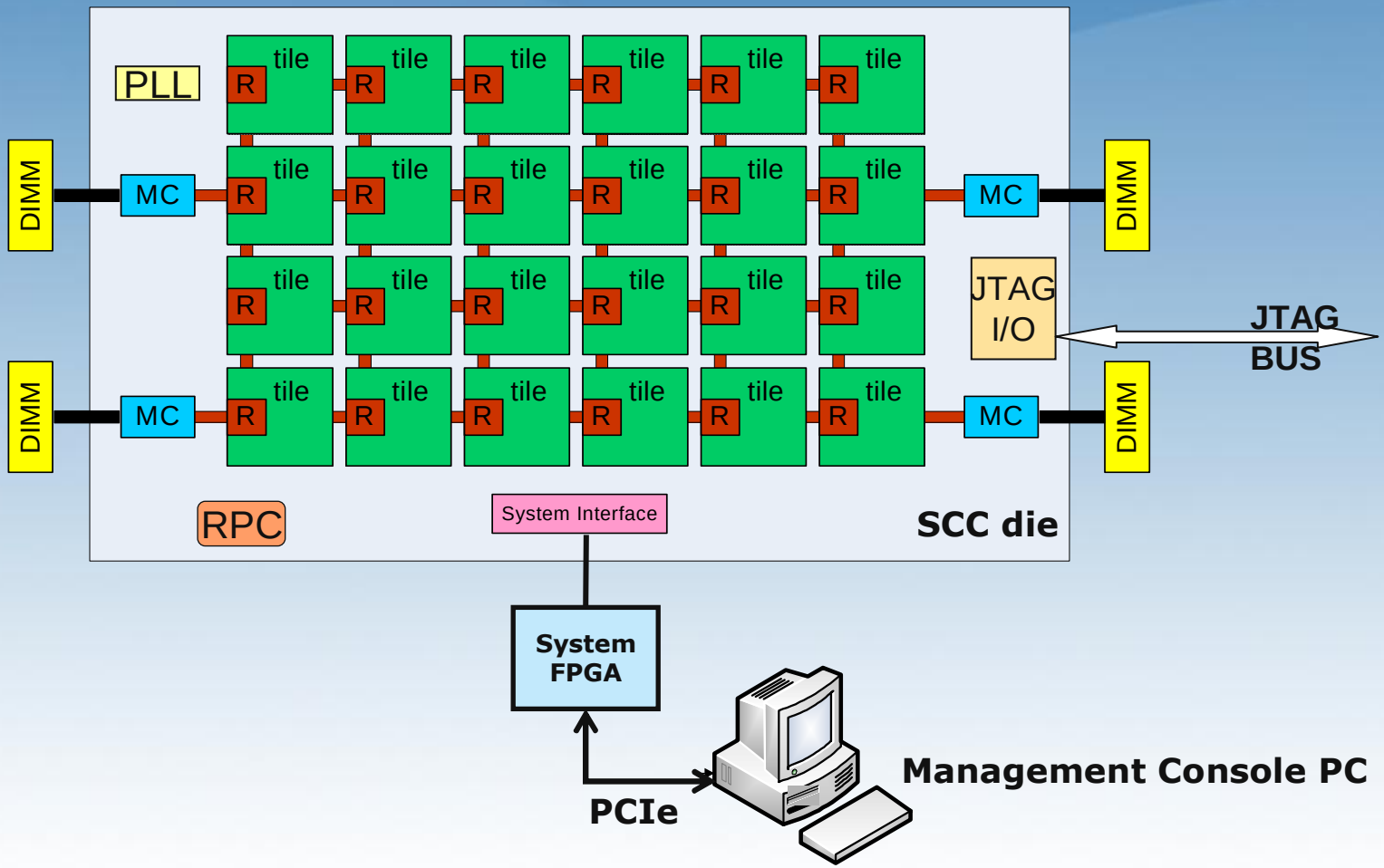
- 16B wide data links + 2B sideband
  - Target frequency: 2GHz
  - Bisection bandwidth: 2 Tb/s
  - Latency: 4 cycles (2ns)
- 2 message classes and 8 virtual channels
  - VC6 for request MCs
  - VC7 for response MCs
- Low power circuit techniques
  - Sleep, clock gating, voltage control, low power RF
  - Low power 5 port crossbar design
- Speculative VC allocation
- Route pre-computation
- Single cycle switch allocation

Router Architecture



Frequency	2GHz @ 1.1V
Latency	4 cycles
Link Width	16 Bytes
Bandwidth	64GB/s per link
Architecture	8 VCs over 2 MCs
Power Consumption	500mW @ 50°C

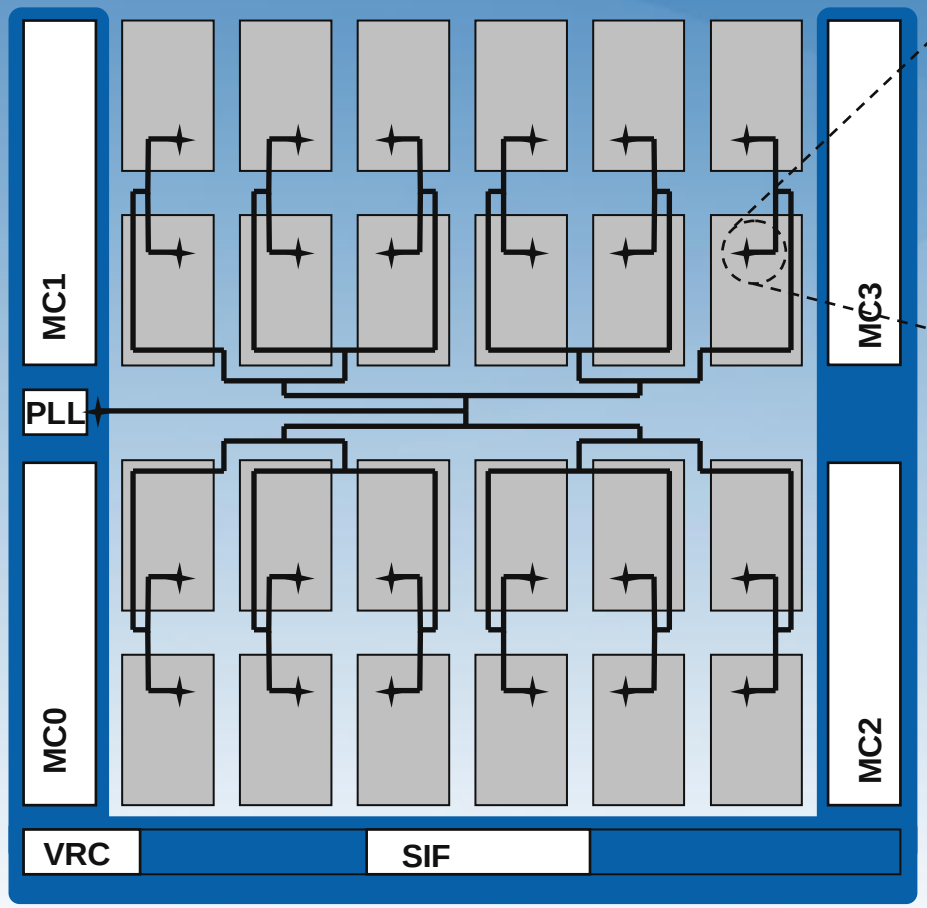
SCC system overview

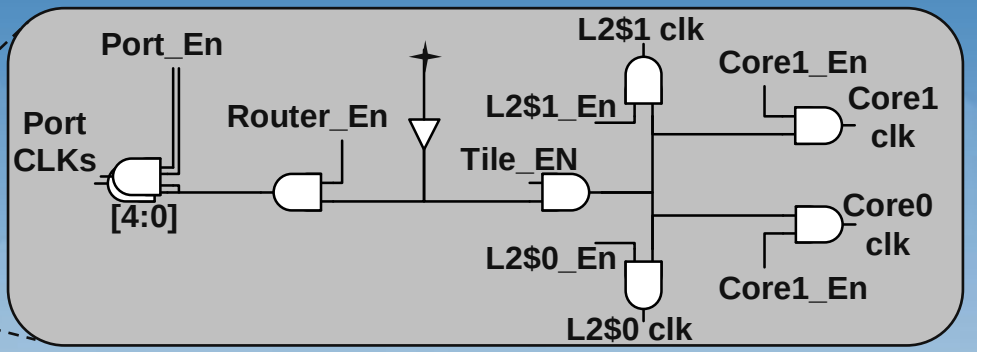


System Interface

- JTAG access to config system while in reset/debug
  - Done on Power Reset from Host
  - Configuring memory controller etc.
  - Reset cores with default configuration
- Management Console PC can use Mem-mapped registers to modify default behavior
  - Configuration and voltage control registers
  - Message passing buffers
  - Memory mapping
- Preload image and reset rather than PC bootstrap
  - BIOS & firmware a work in progress

Clock Distribution





■ Router ▭ Tile

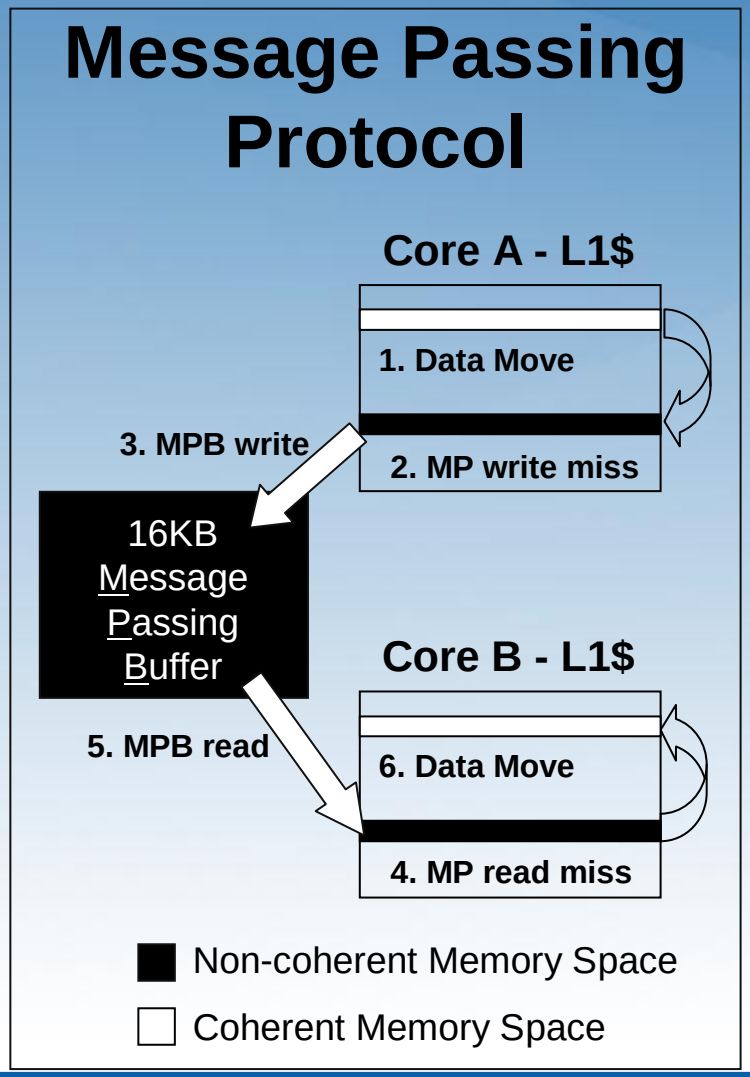
✦ Clock Gating

- Balanced H-tree clock distribution
- Designed to provide 4GHz clock to tile entry points
- Simulated skew for adjacent tiles – 5ps
- Cross die skew irrelevant

Message Passing on RC

- Message passing is done through shared memory space
- Two classes of shared memory:
  - Off-die, DRAM: Uncacheable shared memory … results in high latency message passing
  - On-die, message passing buffers (MPB) … low latency message passing
- MPB performance
  - Message Passing Buffers see a 15x improved latency as compared to off die DDR3-800

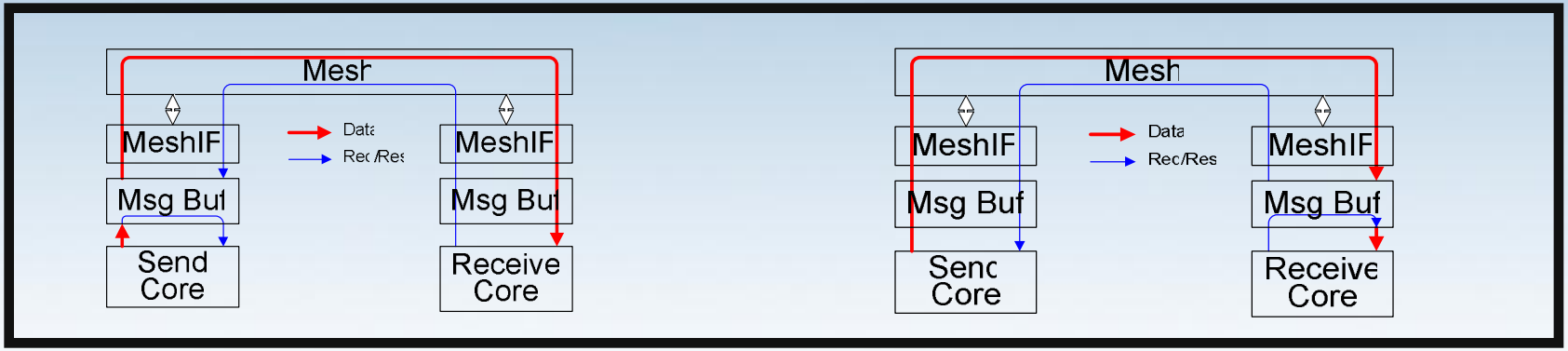
Message Passing Protocol



- Cores communicate through small fast messages
  - L1 to L1 data transfers
  - New Message Passing Data Type (MPDT)
- Message passing Buffer (MPB) – 16KB
  - 1 MPB per tile for 384KB of on-die shared memory
  - MPB size coincides with L1 caches

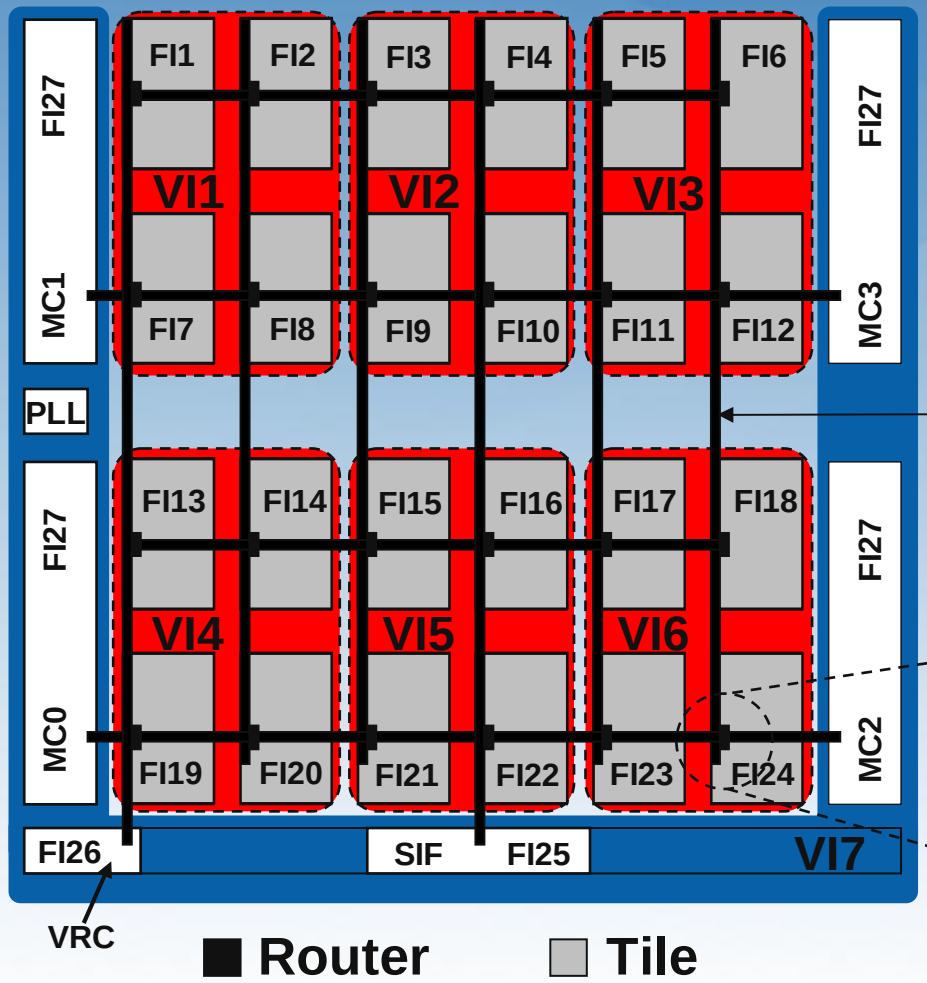
Dedicated Message Buffers

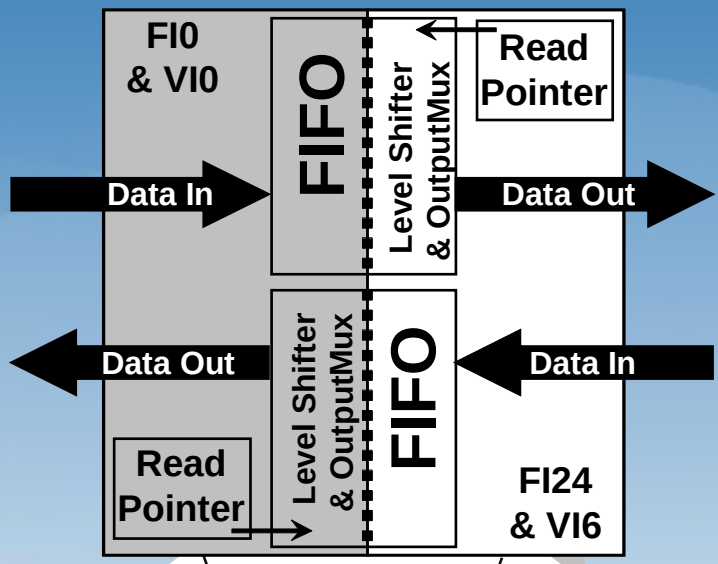
- Messages can be read from / written to one of three locations.
  - A message buffer locally in a core's tile.
  - A message buffer remotely in another tile
  - Off die to main memory
- We believe a remote write, local read has the best performance

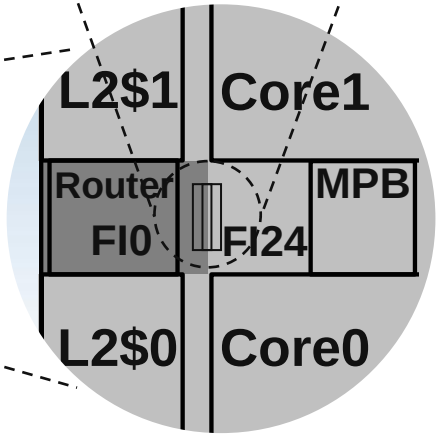


Local write, remote read **Remote write, local read**

Voltage and Frequency islands



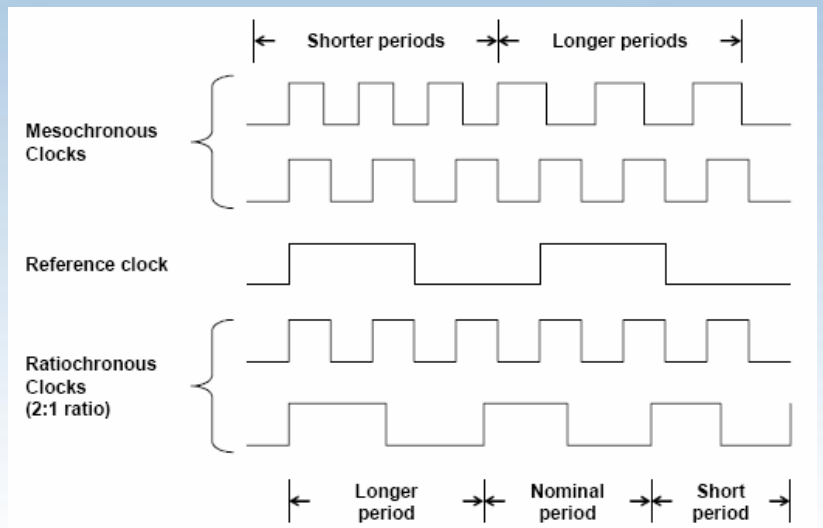


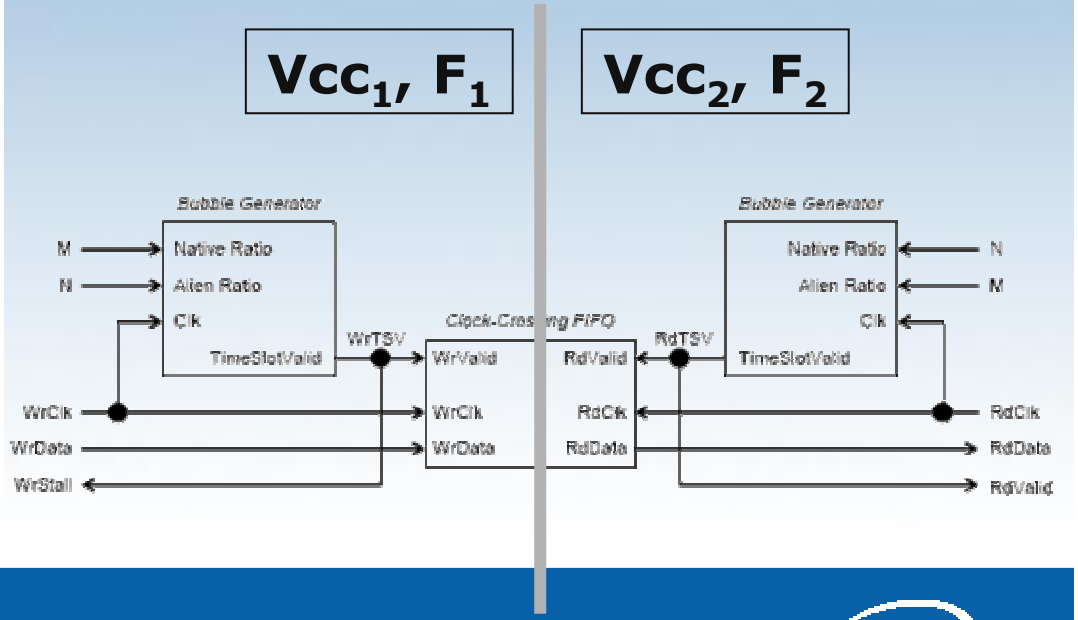


28 Frequency Islands (FI) 8 Voltage Islands (VI)

SCC Clock Crossing FIFO (CCF)

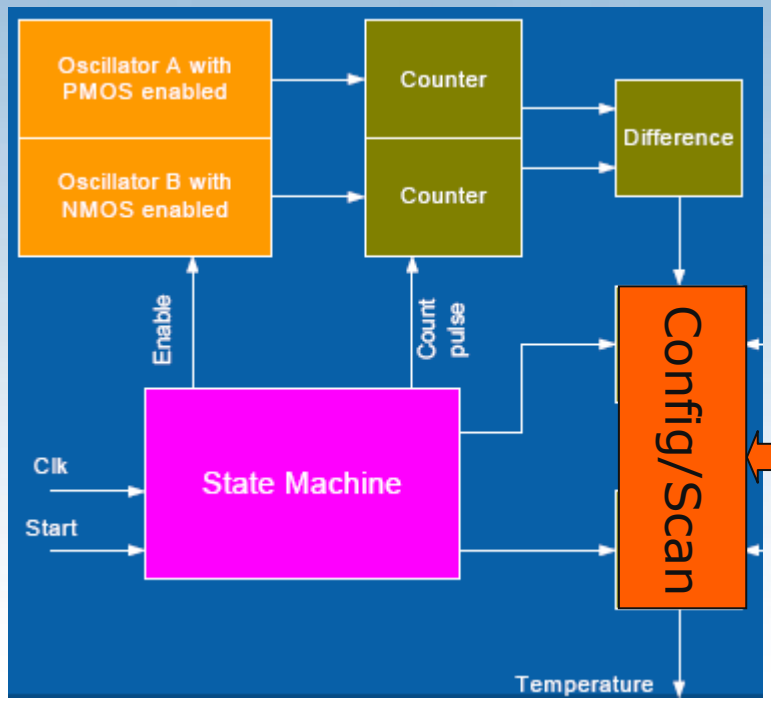
- 6 entry deep FIFO, 144-bits wide
- Built-in Voltage translation: 1:N ratios and pointer separation scanned in
- Key Benefit: independent mesh & tile frequency





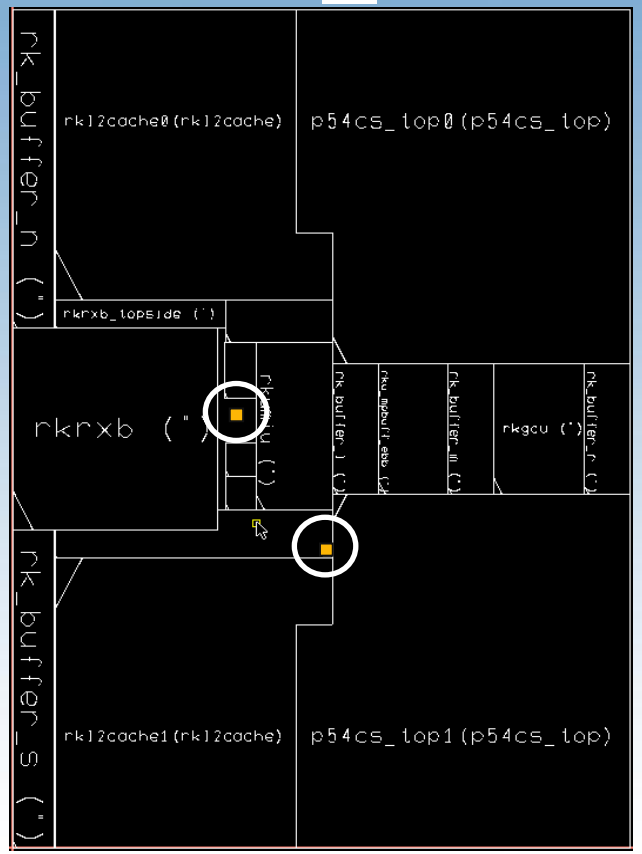
Closed Loop Thermal Management

- Network of digital temperature sensors
  - 2 per Tile, 48 total
  - Programmable 13-bit counters
  - Outputs written to config registers
    - > Readable by any P54c core for DVFS

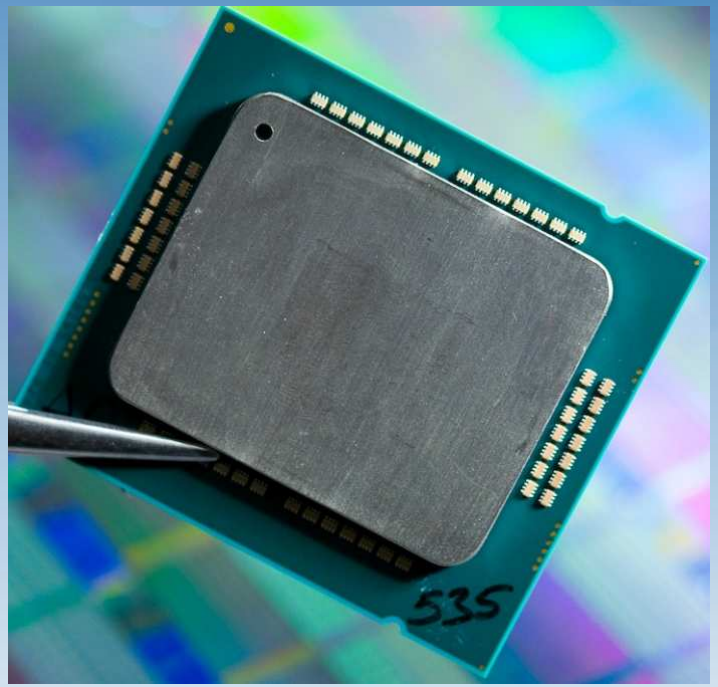


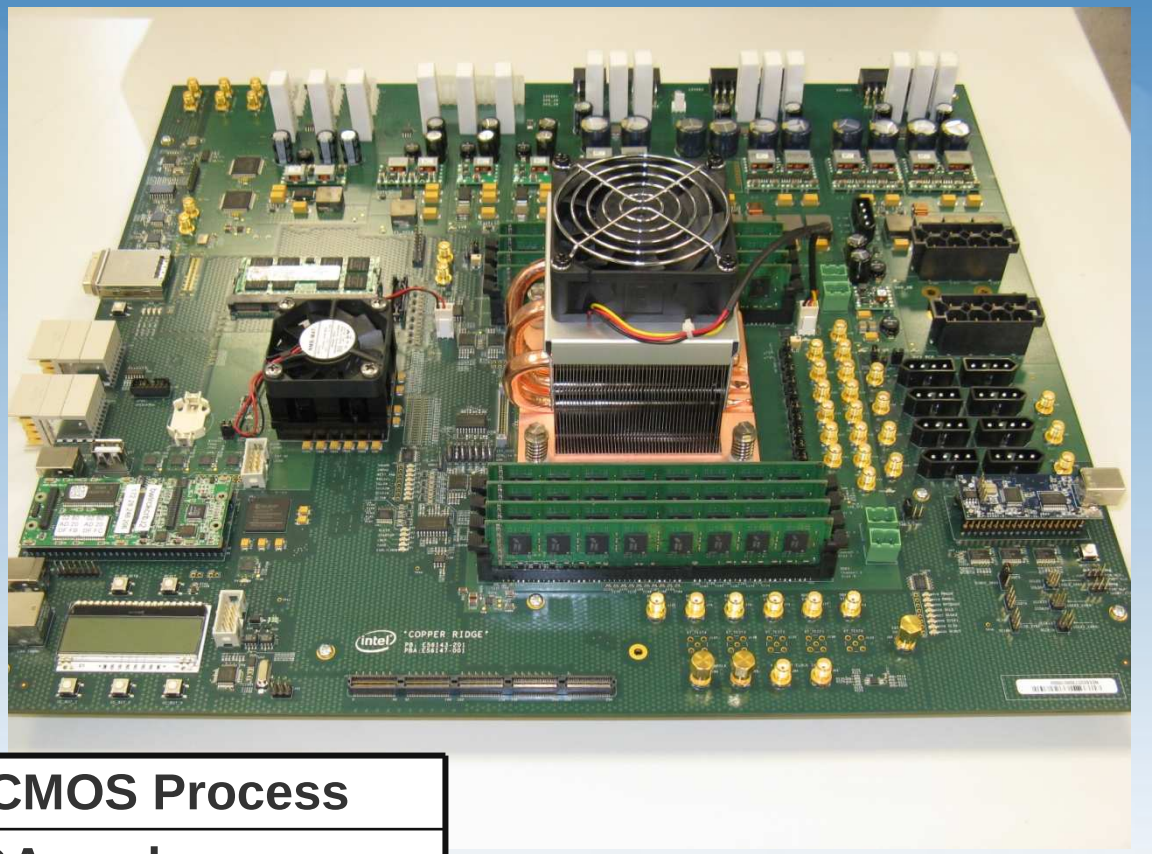
To Config Regs

SCC_TILE



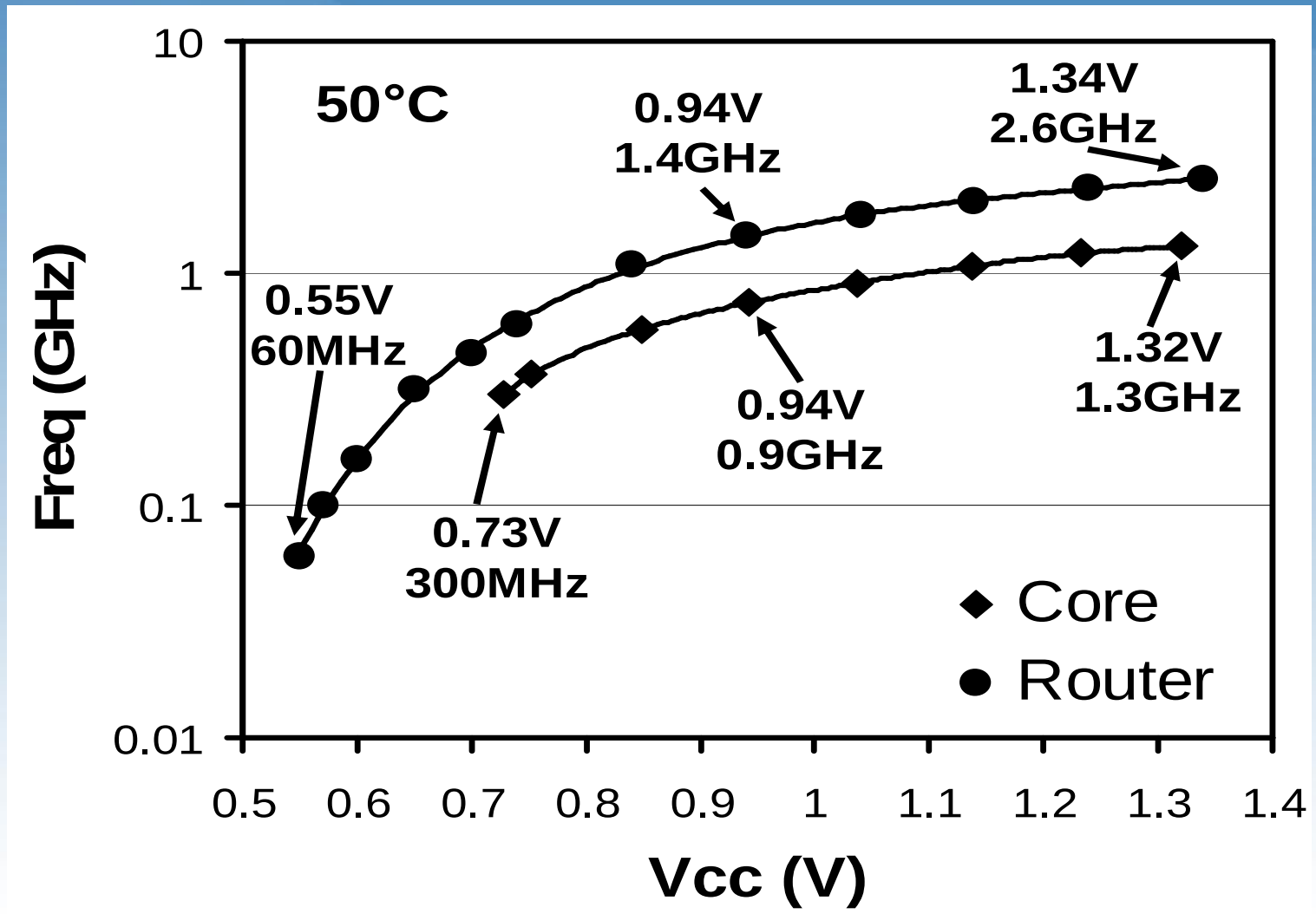
Package and Test Board



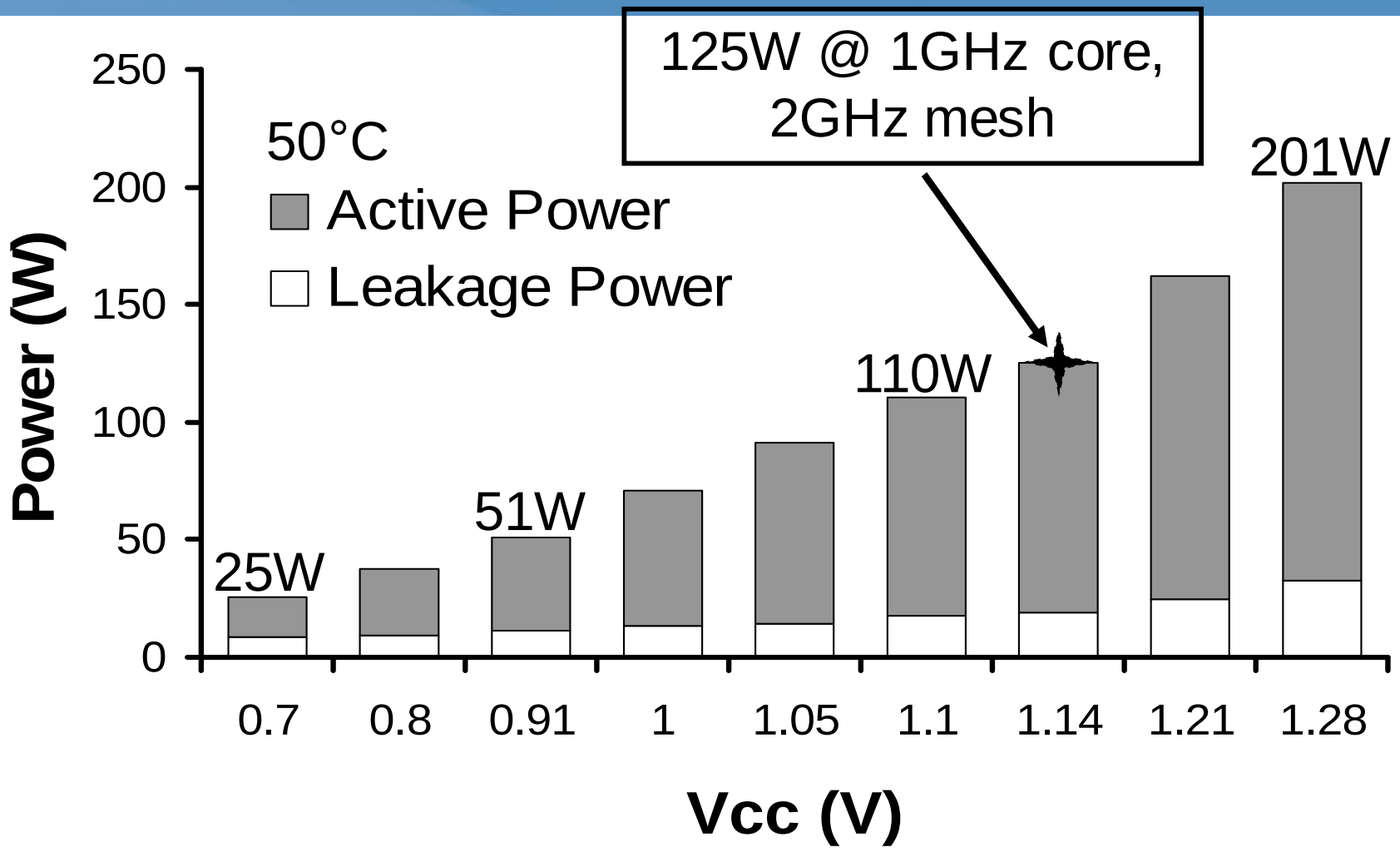


Technology	45nm Hi-K CMOS Process
Package	1567 pin LGA package
	14 layers (5-4-5)
Signals	970 pins

Core & Router Fmax

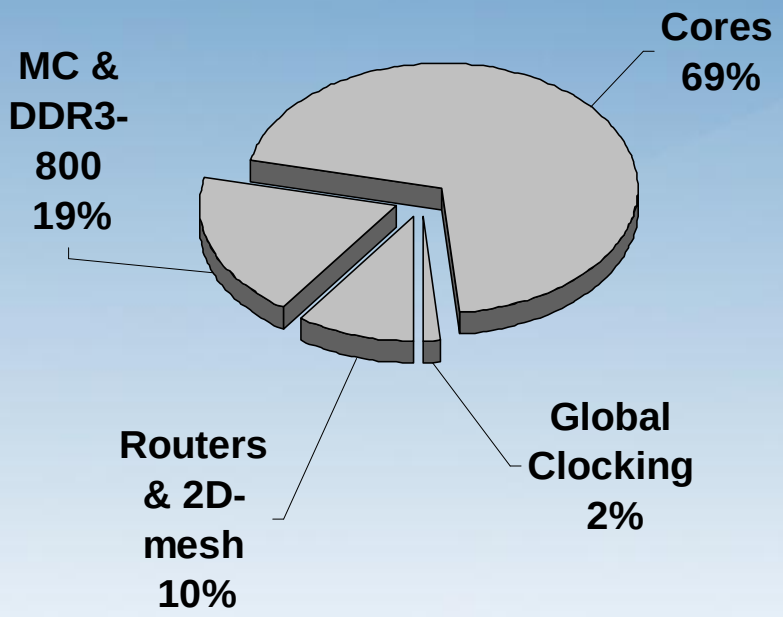


Measured full chip power



Power breakdown

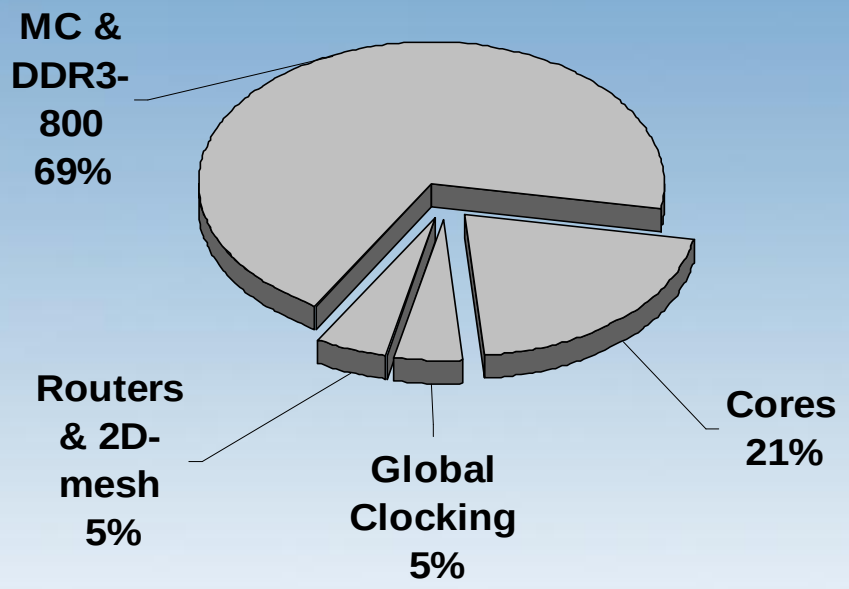
Full Power Breakdown
Total -125.3W



Clocking: 1.9W Routers: 12.1W
Cores: 87.7W MCs: 23.6W

Cores-1GHz, Mesh-2GHz, 1.14V, 50°C

Low Power Breakdown
Total - 24.7W



Clocking: 1.2W Routers: 1.2W
Cores: 5.1W MCs: 17.2W

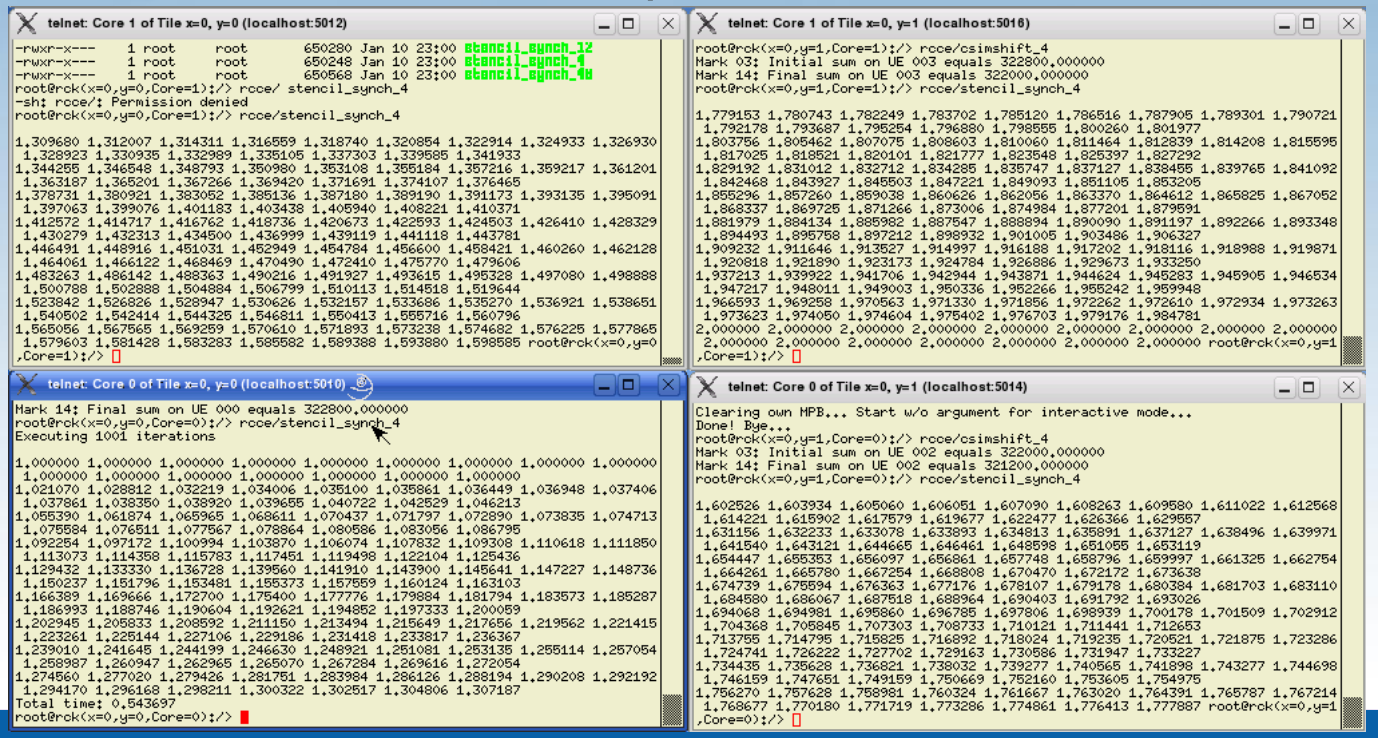
Cores-125MHz, Mesh-250MHz, 0.7V, 50°C

Linux

A small Linux build was created targeted st the SCC specific features. The Linux memory driver was modified to enable Linux control of the on-die message passing buffers. Included in this Linux build is a TCP/IP driver for the 2-D mesh, connecting the host-PC and all 48 cores at the software layer.

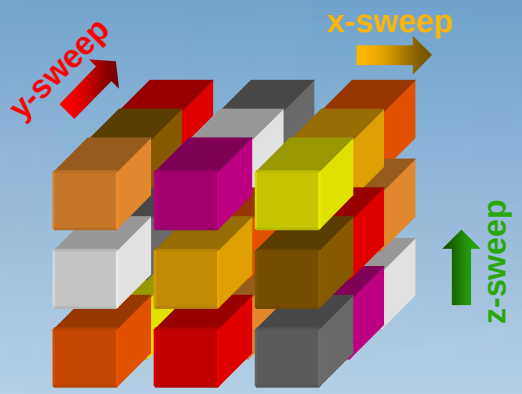
Also, a TTY driver is included to allow the 48 cores to perform IO commands via memory mapped IO.

This enables regular xterm type connections as shown.



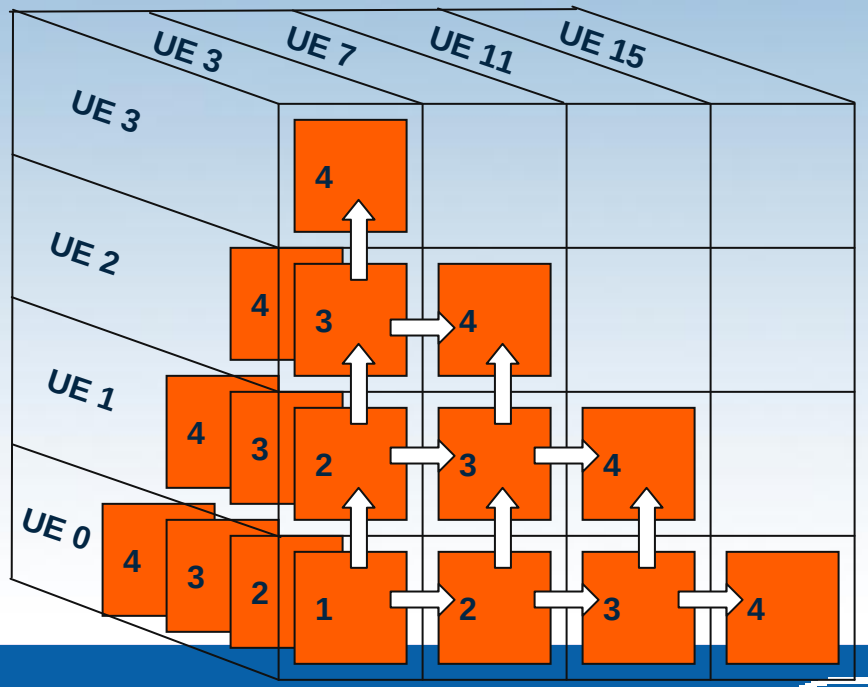
Linpack and NAS Parallel benchmarks

1. **Linpack (HPL): solve dense system of linear equations**
   - Synchronous comm. with "MPI wrappers" to simplify porting



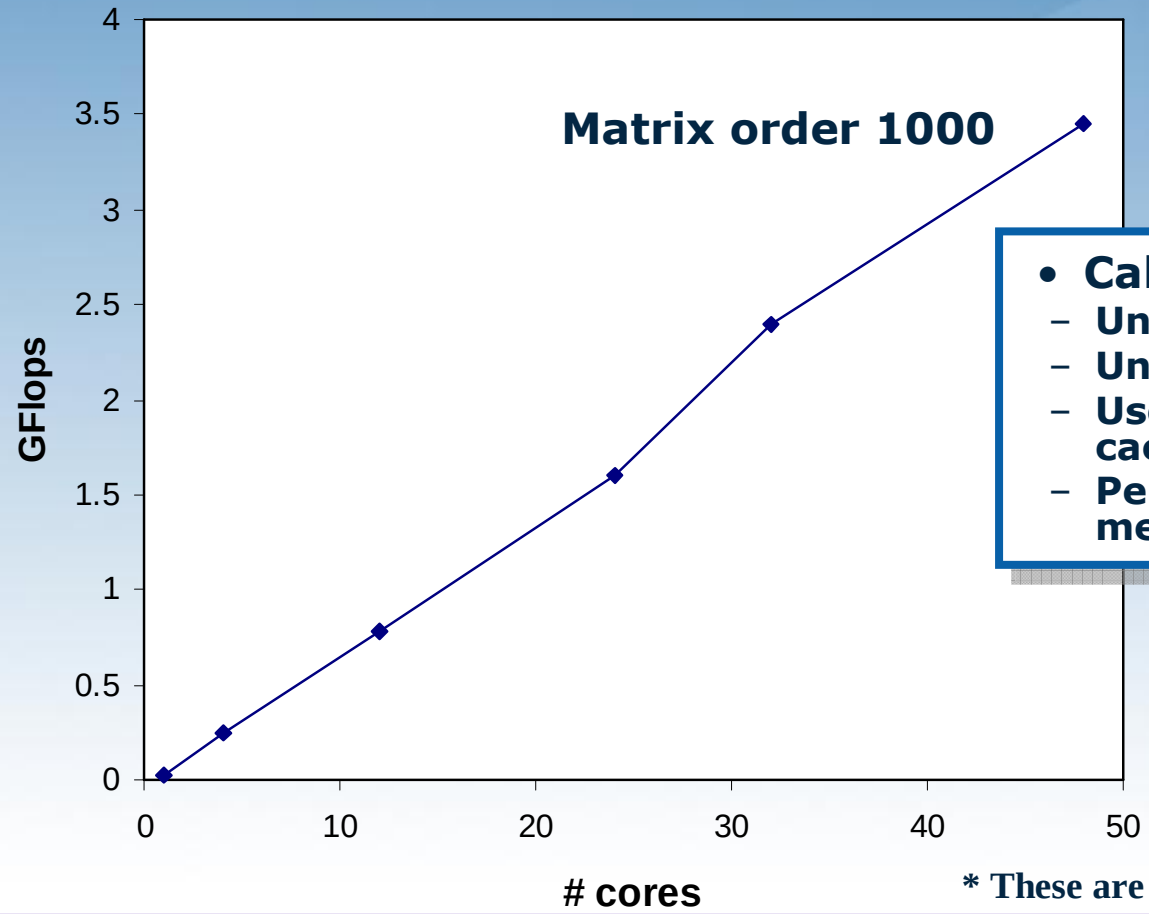
2. **BT: Multipartition decomposition**
   - Each core owns multiple blocks (3 in this case)
   - update all blocks in plane of 3x3 blocks
   - send data to neighbor blocks in next plane
   - update next plane of 3x3 blocks

3. **LU: Pencil decomposition**
   Define 2D-pipeline process
   - await data (bottom+left)
   - compute new tile
   - send data (top+right)



Linpack, on the Linux SCC platform

- **Linpack (HPL)* strong scaling results:**
  - **GFLOPS vs. # of cores for a fixed size problem (1000).**
  - **This is a tough test … scaling is easier for large problems.**

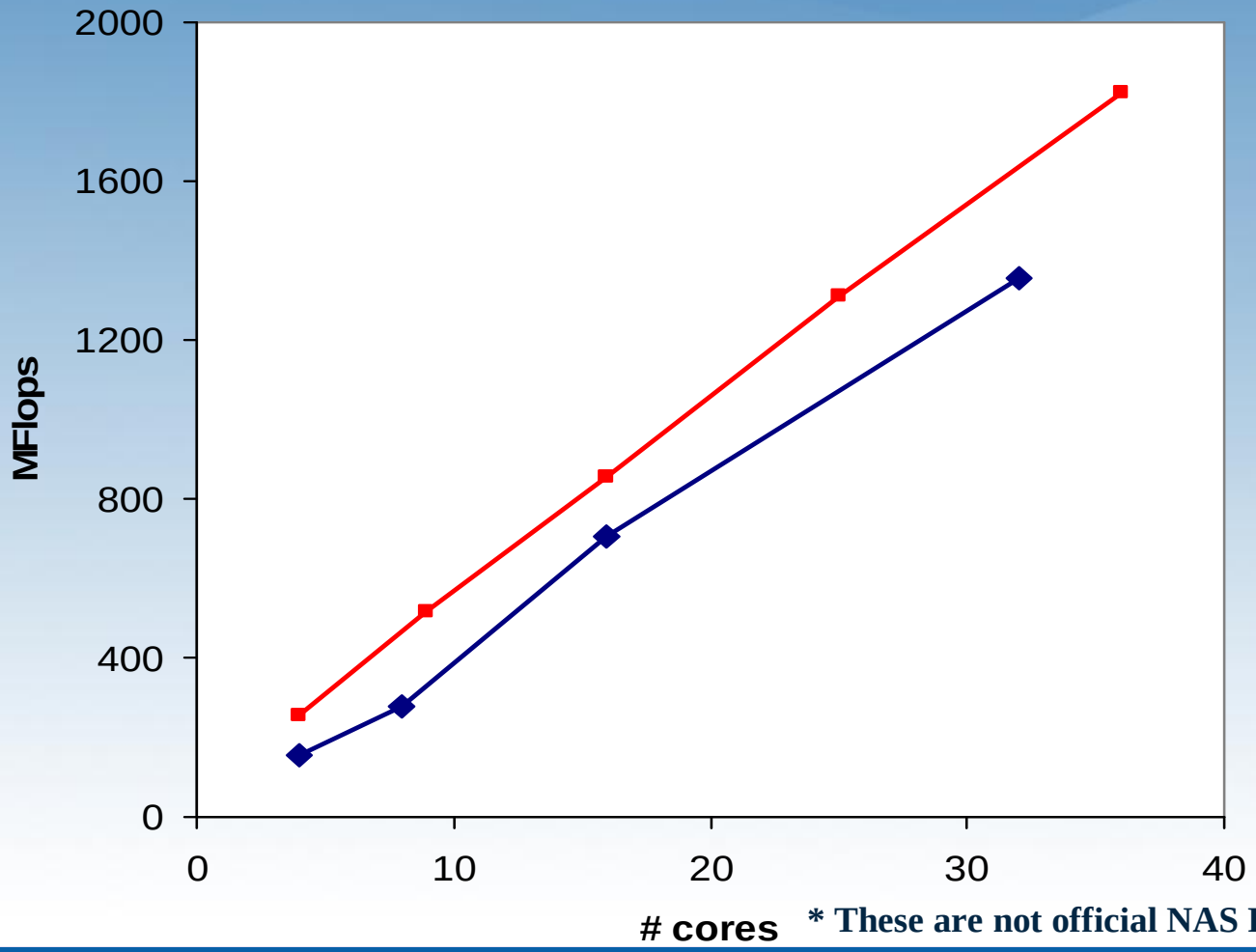


- **Calculation Details:**
  - **Un-optimized C-BLAS**
  - **Un-optimized block size (4x4)**
  - **Used latency-optimized whole cache line flags**
  - **Performance dropped ~10% with memory optimized 1-bit flags**

*** These are not official LINPACK benchmark results.**

LU/BT NAS Parallel Benchmarks, SCC

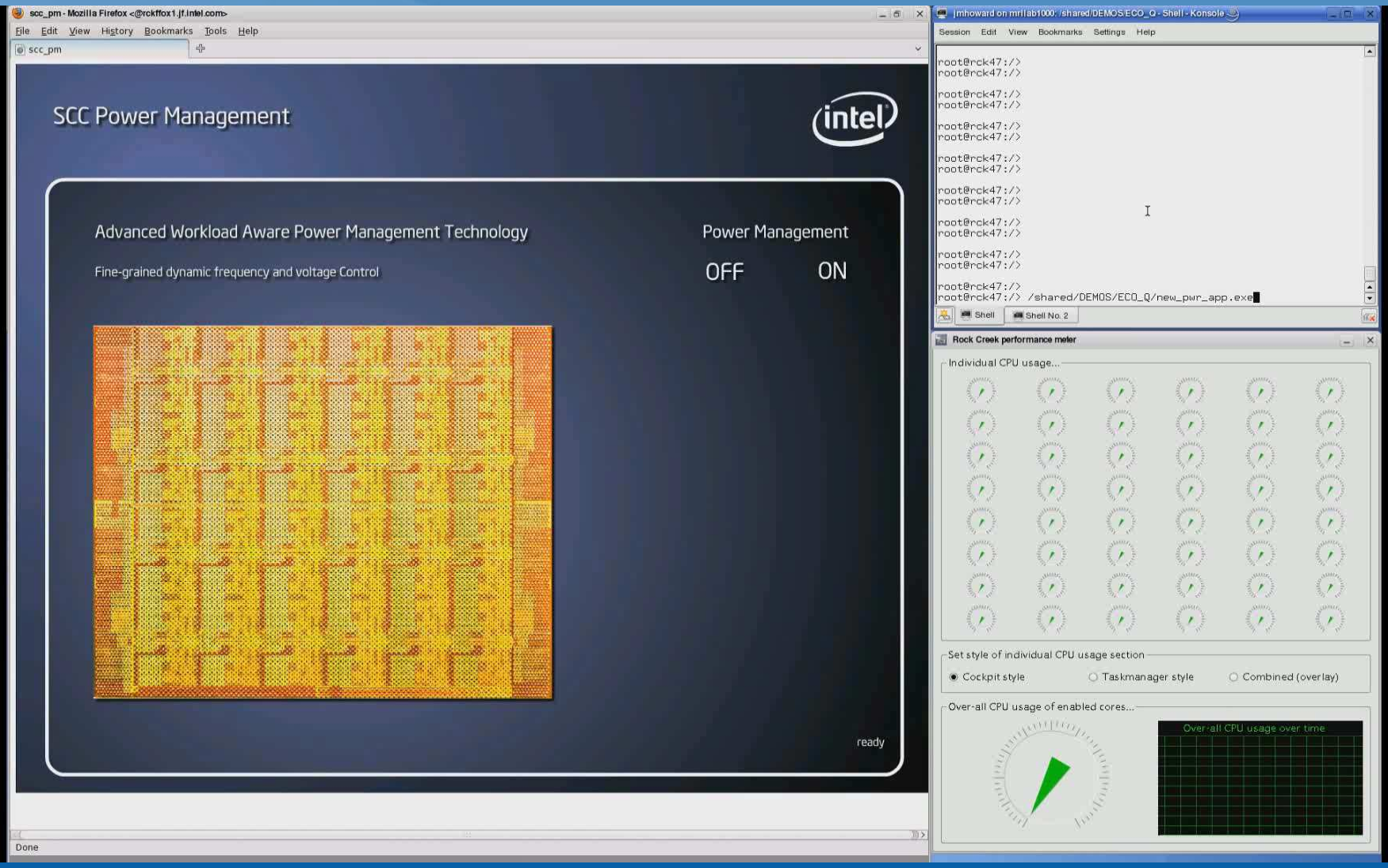
Problem size: Class A, 64 x 64 x 64 grid*



- **Using latency optimized, whole cache line flags**

*** These are not official NAS Parallel benchmark results.**

Power Management Demo



Summary

- A 48 IA-32 core processor in 45nm CMOS
  - Second generation 2D-mesh network
  - 4 DDR3 channels in a 6×4
  - Highest level of IA-32 integration
- New message passing HW for increased performance
  - 384KB of on-die shared memory
  - Message passing memory type
- Power management employs 8VIs and 28FIs for DVFS
- Chip dissipates between 25W and 125W as performance scales
  - 25W at 0.7, 125MHz core, 250MHz mesh and 50°C
  - 125W at 1.14V, 1GHz core, 2GHz mesh and 50°C