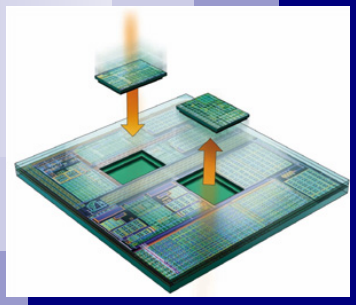
Presentation at MPSoC 2008
This work was supported in part by NSF Award **0702617**

June 23, 2008



Network-on-Chip for 3D Architectures

Vijaykrishnan Narayanan

Collaborators:
C. Nicopoulos, R. Das, S. Eachempati, A. Mishra, J. Kim, Y. Xie, D. Park, C.R. Das

The Pennsylvania State University
Department of Computer Science & Engineering
Microsystems Design Lab (www.cse.psu.edu/~mdl)

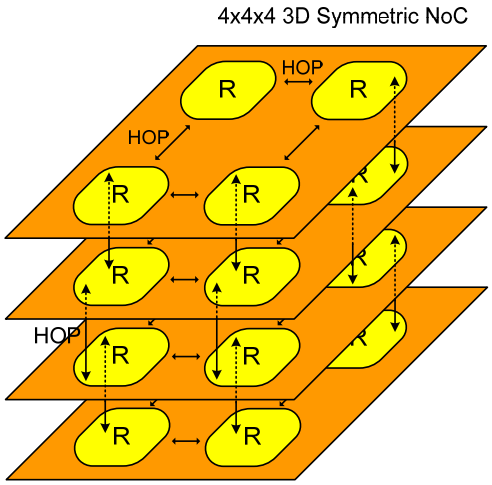
Exploring 3D NoC Architectures

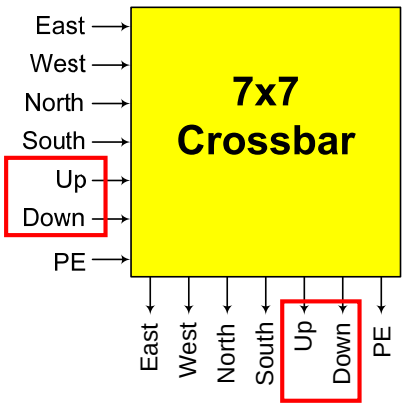
Investigation of several *design options for 3D NoCs*
(Specifically, focus on the *inter-strata communication*)

- **A 3D *Symmetric* NoC Architecture**
- **A 3D *NoC-Bus Hybrid* Architecture**
- **A *Full 3D* NoC Router**
- **3D Dimensionally Decomposed (*3D DimDe*) Router**
- **Multilayer Router**

The *3D Symmetric* NoC Architecture

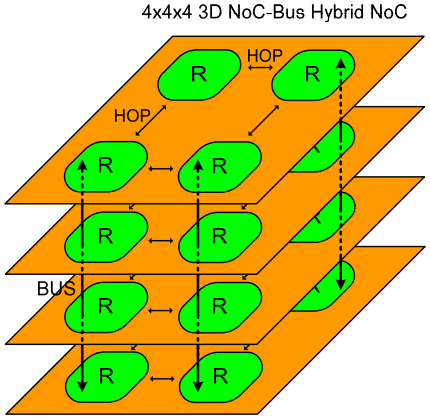
- Simplest extension to the generic 2D NoC router to facilitate a 3D layout:
  - **Hop-by-Hop Traversal** (2D Crossbar)

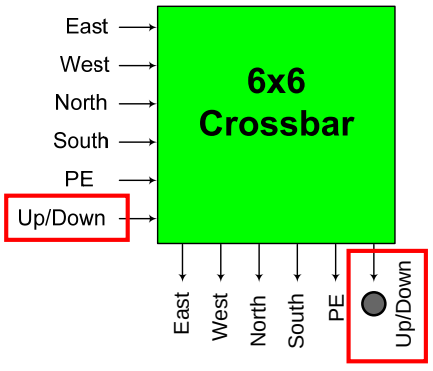


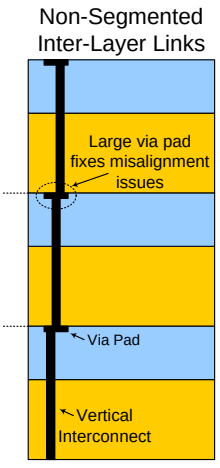


The *3D NoC/Bus Hybrid* Architecture

- NoC fabric is hybridized with a ***bus link*** in the vertical dimension
- Very small inter-strata distance
  - ***Single-Hop Communication*** is feasible





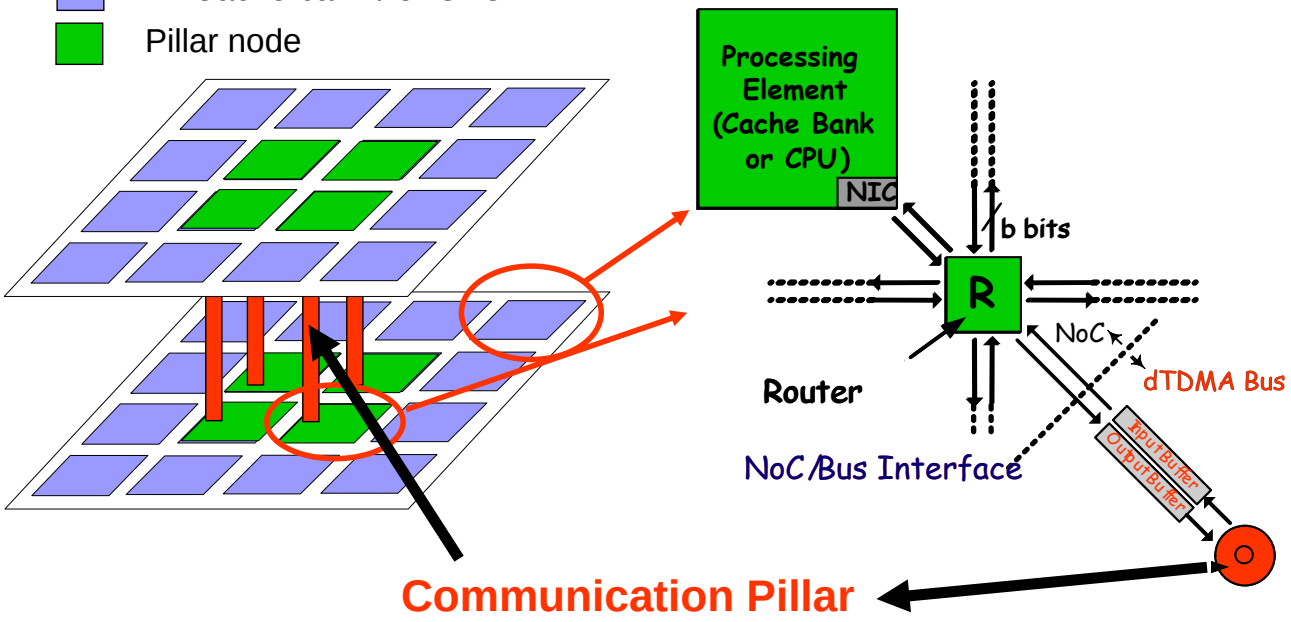


Inter-Layer Via Structure

The 3D Network-in-Memory (NetInMem)

L2 Cache bank / or CPU

Pillar node

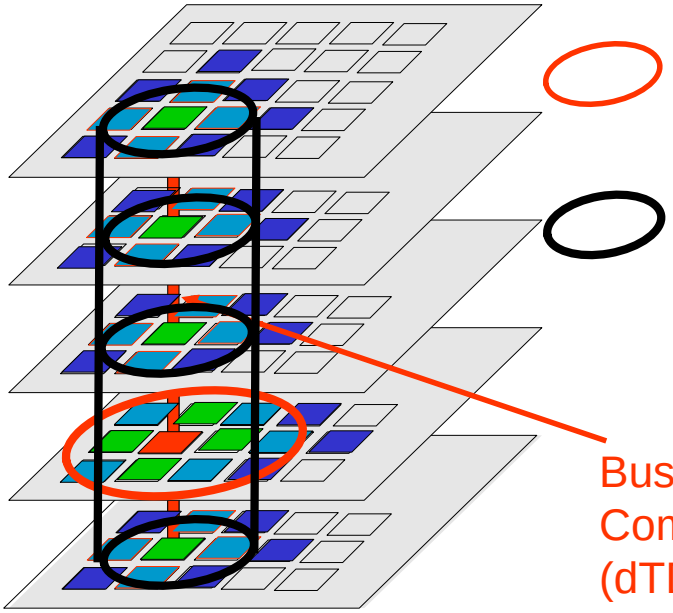


Communication Pillar

dTDMA Bus (Dynamic Time Division Multiple Access)

3D Benefit: *Increased Locality*

- CPU
- Nodes within 1 hop
- Nodes within 2 hops
- Nodes within 3 hops

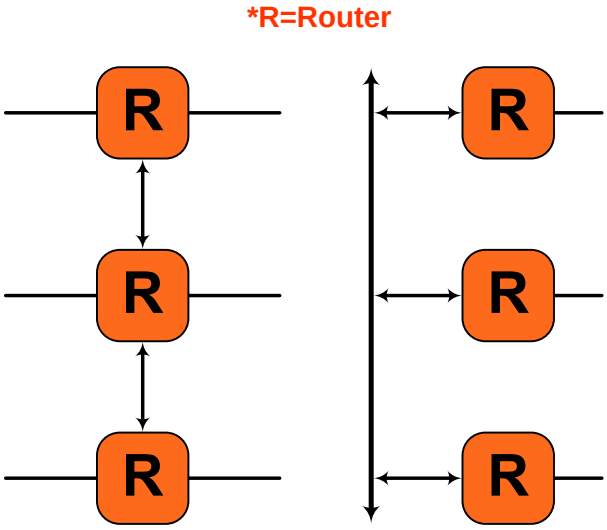


2D vicinity

3D vicinity

Bus-based Inter-Layer Communication (dTDMA Bus Pillar)

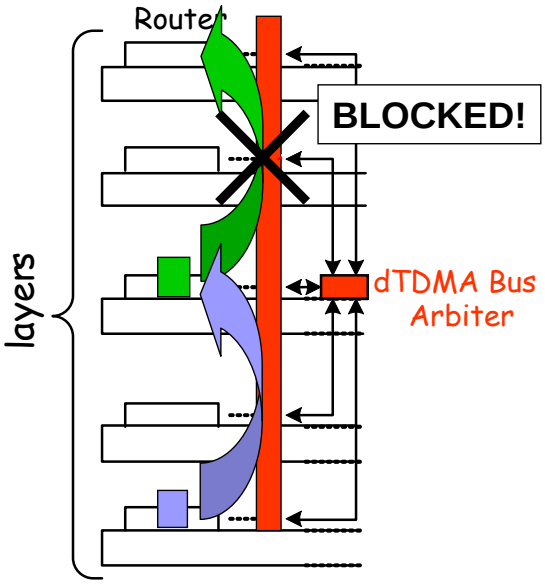
Problem: Contention for the Bus



Hop-by-Hop

Single-Hop Bus

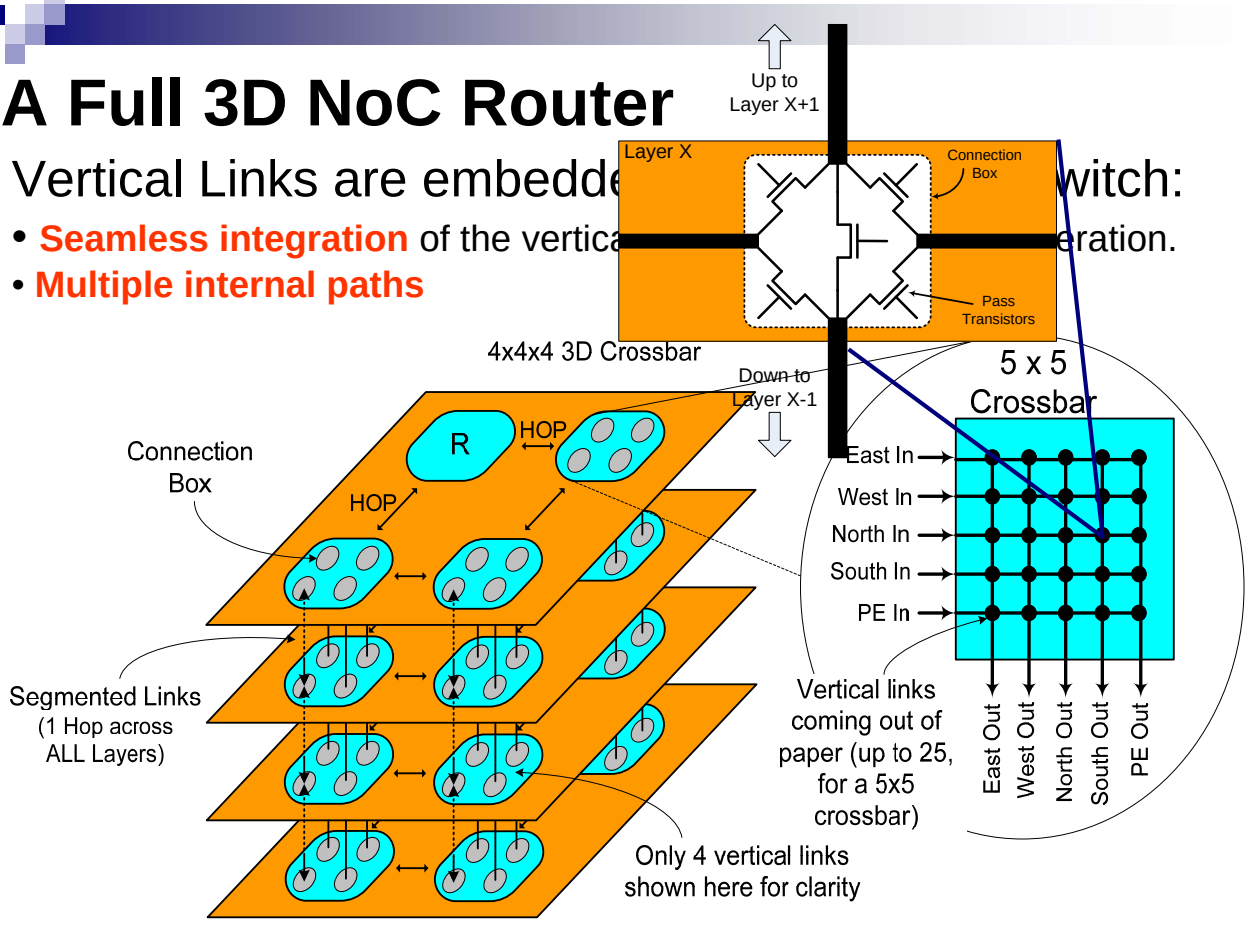
[NetInMem ISCA-2006]



A Full 3D NoC Router

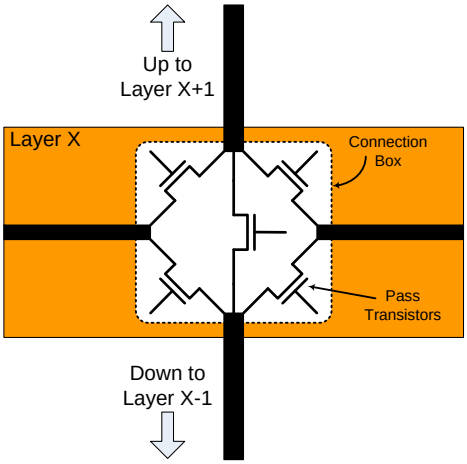
Vertical Links are embedde witch:

- **Seamless integration** of the vertica eration.
- **Multiple internal paths**

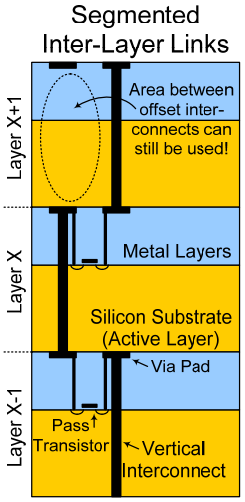


Inter-Layer Via Structure in a 3D Crossbar Technology

3D connection box can facilitate linkage between vertical and horizontal channels

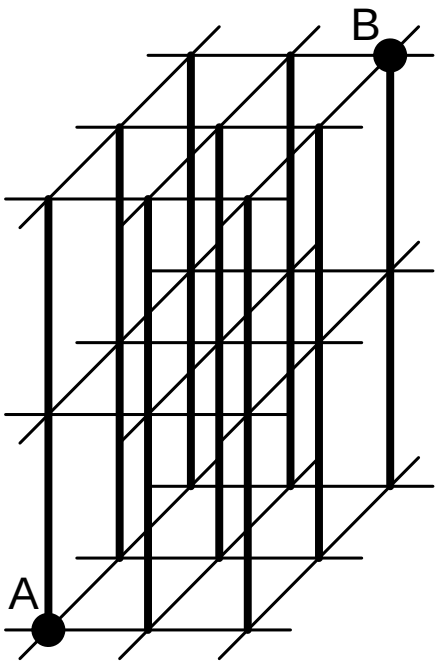


Connection Box



Inter-Layer Via Layout (Vertical Link Segmentation)

Daunting Path Diversity!



$$k = \begin{pmatrix} \Delta_x + \Delta_y + \Delta_z \\ \Delta_x \end{pmatrix} \begin{pmatrix} \Delta_y + \Delta_z \\ \Delta_y \end{pmatrix} = \frac{(\Delta_x + \Delta_y + \Delta_z)!}{\Delta_x!\Delta_y!\Delta_z!}$$

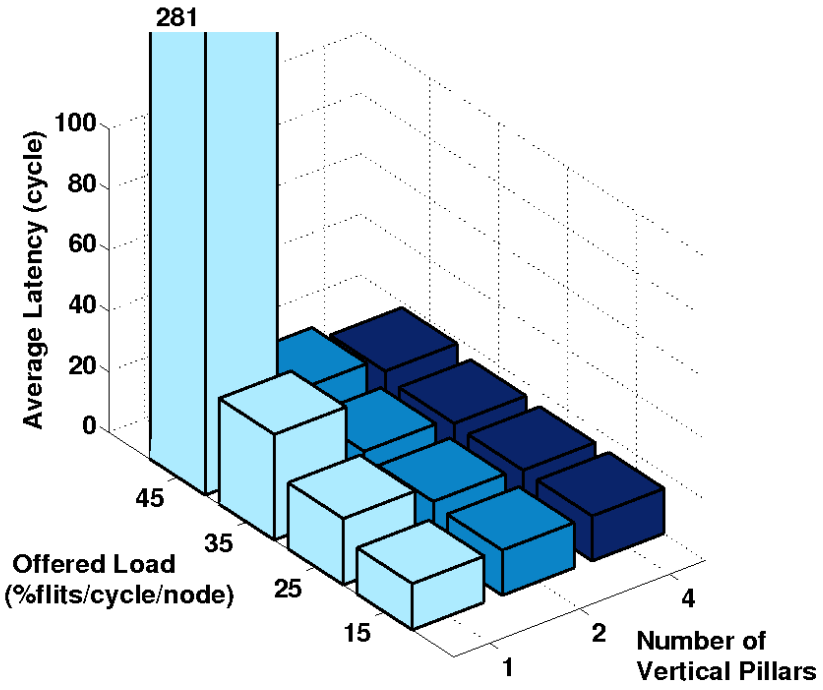
k = Number of **Minimal Paths between A and B**

Delta$_x$, ***Delta$_y$***, ***Delta$_z$*** = **Number of hops** separating A and B in x, y, and z dimensions, respectively

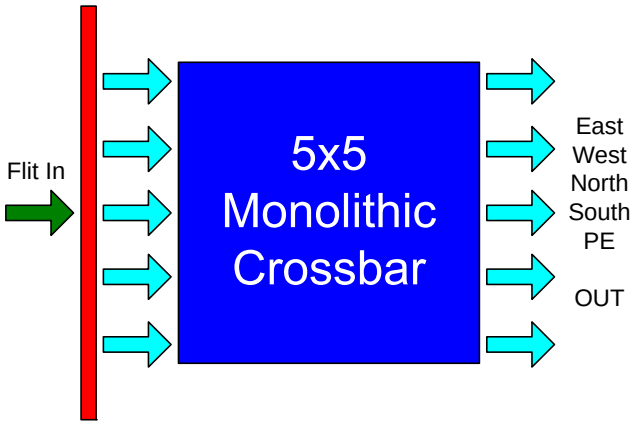
3x3x3 Crossbar (shown at left), ***k = 90***
4x4x4 Crossbar, ***k = 1680***
…

…and these are just the MINIMAL paths!!!

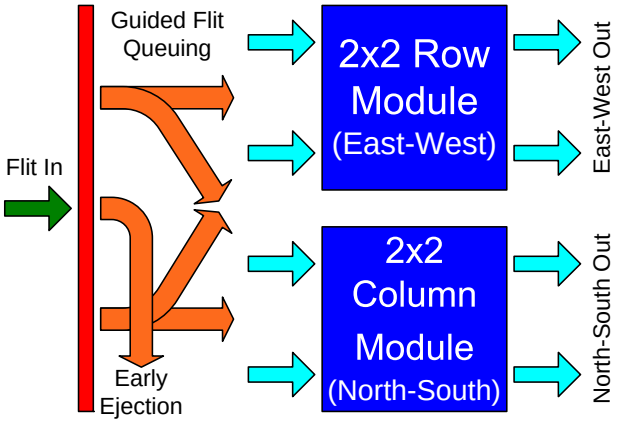
Impact of the *Number of Vertical Bundles* on Performance



Dimensionally Decomposed Routers

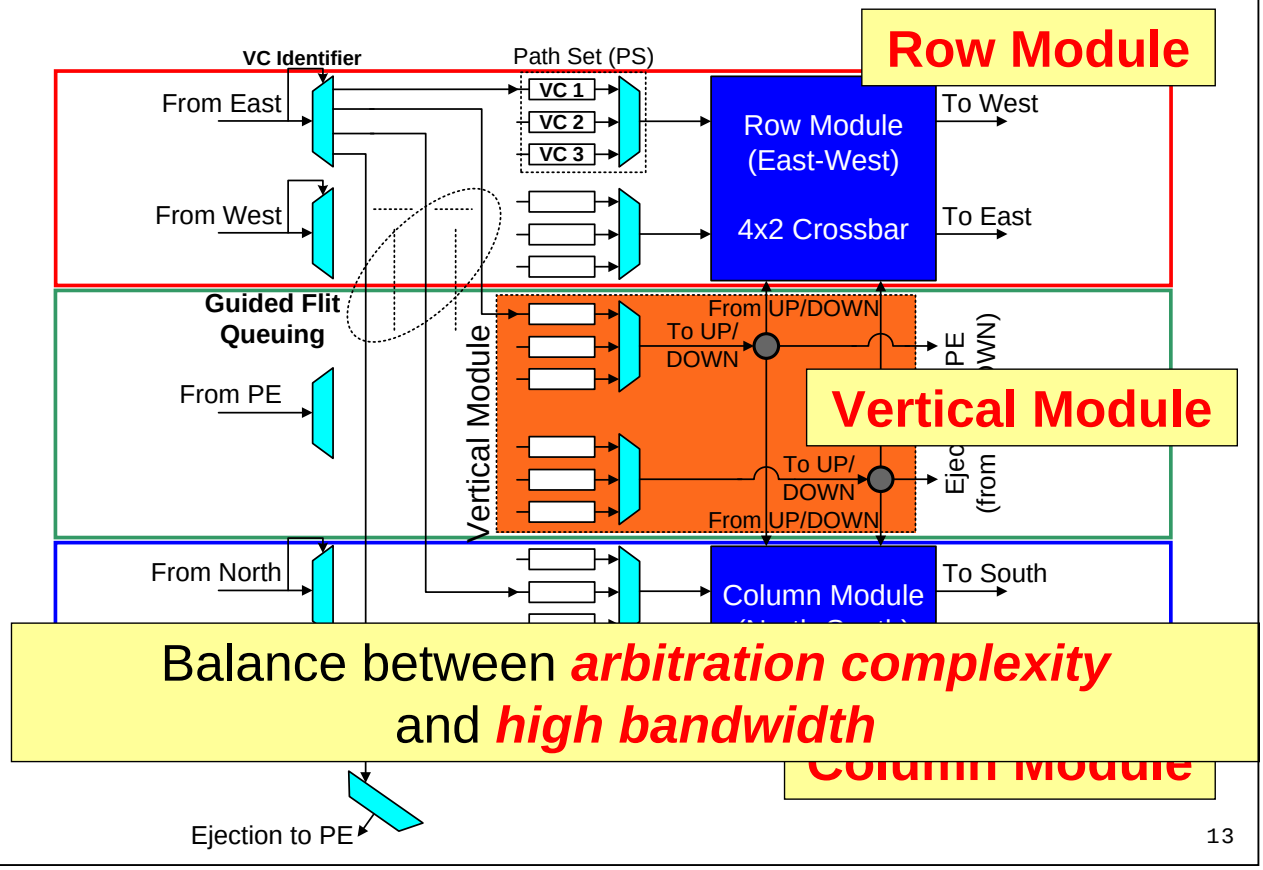


A Conventional NoC Router



The Row-Column (RoCo) Decoupled Router [ISCA-06]

Partitioned Virtual Channels

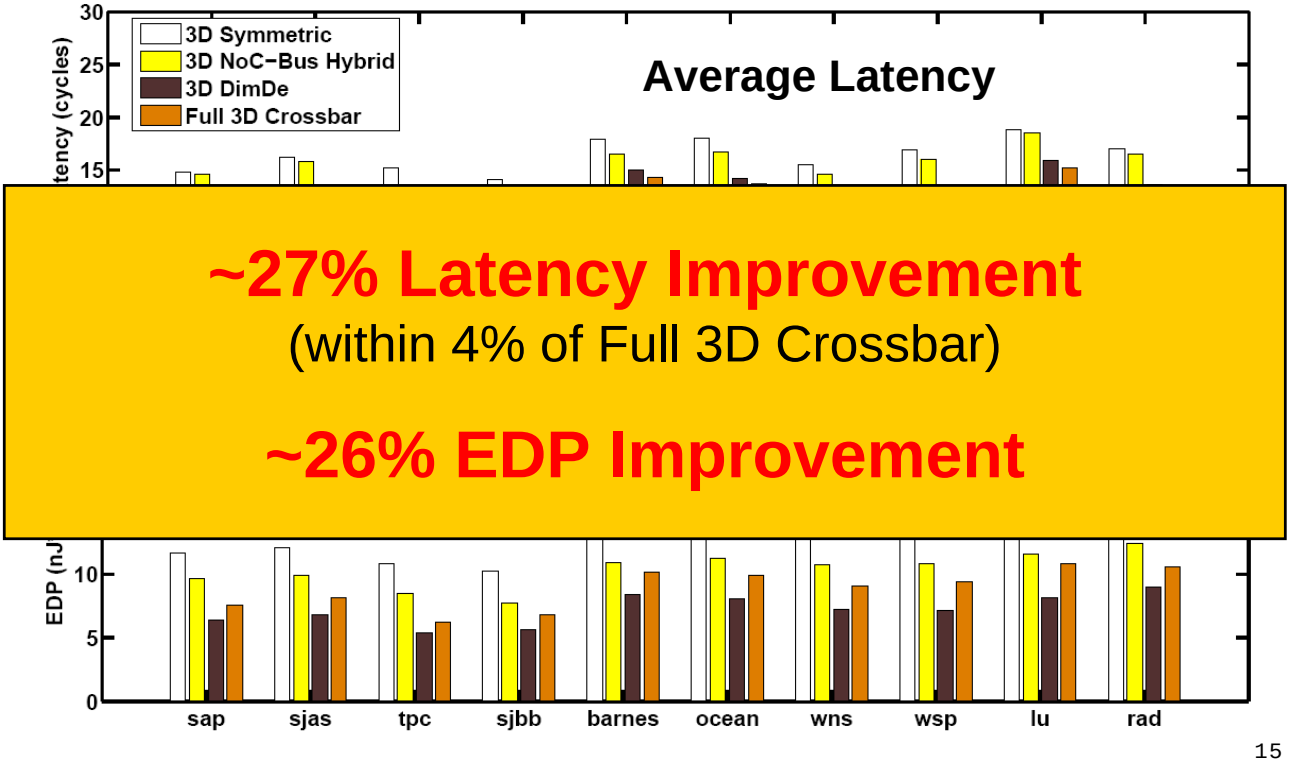


Balance between ***arbitration complexity*** and ***high bandwidth***

Performance Results w/ ***Real*** Workloads

32 L2 Cache Bank Nodes, 512 KB each, 16 MB TOTAL L2
8 Sun UltraSPARC III CPU Nodes

SPLASH Scientific Benchmarks
TPC-C, SAP, SJBB, SJAS Commercial Benchmarks

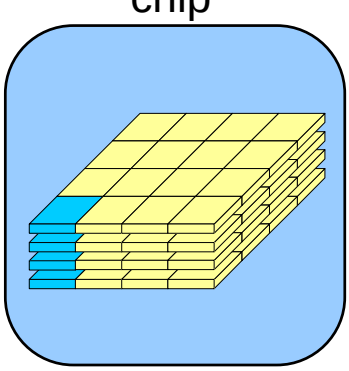


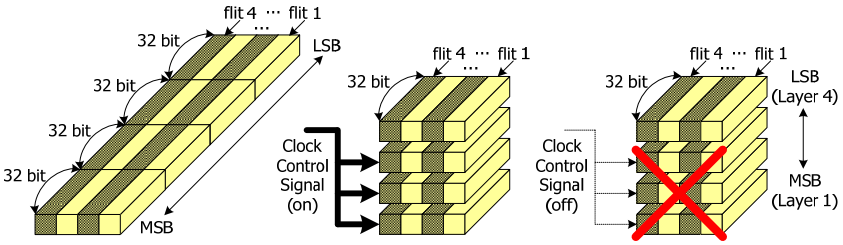
~27% Latency Improvement
(within 4% of Full 3D Crossbar)

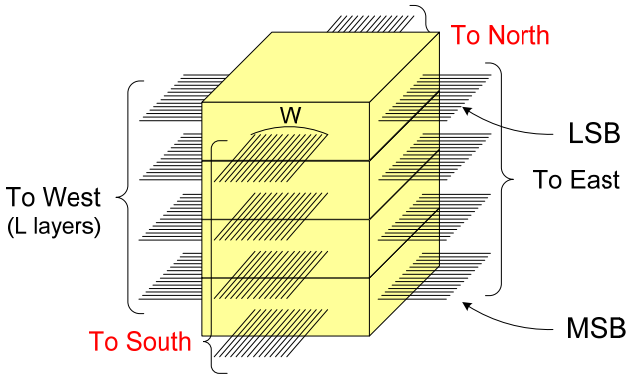
~26% EDP Improvement

Multi-layer On-Chip Interconnect Router Architecture (MIRA)

- 3D Multi-layered (3DM) router architecture designed to span across the multiple layers of a 3D chip

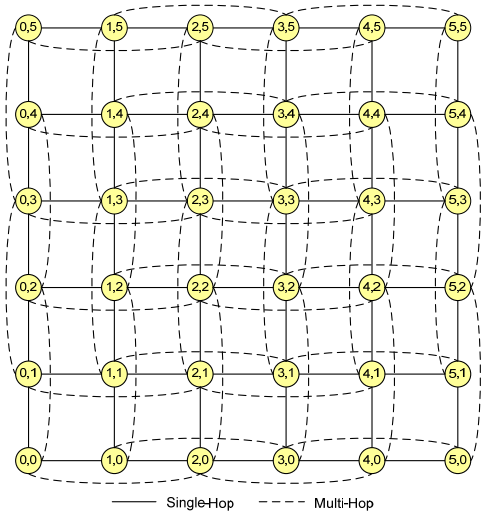






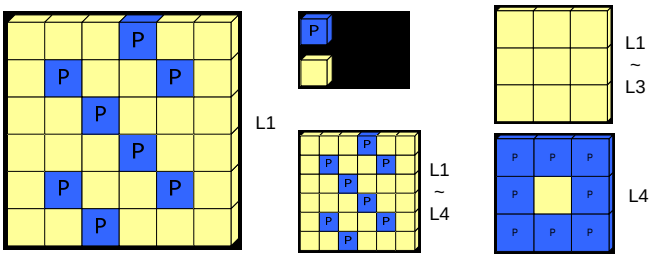
MIRA: Extension (3DM-E)

- Exploit additional bandwidth for additional physical express channels



MIRA: Simulation Setup

- Node Layout:
  - Total 36 cores:
    - 8 CPU cores
    - 28 L2 cache cores
  - Core: Sun Niagara
  - Cache: 512KB each (14MB total)



(a) 2DB (b) 3DM-E (c) 3DB

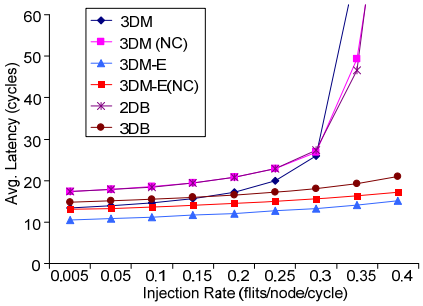
[Node Layouts for 36 cores]

- Cache Configuration (benchmarks)

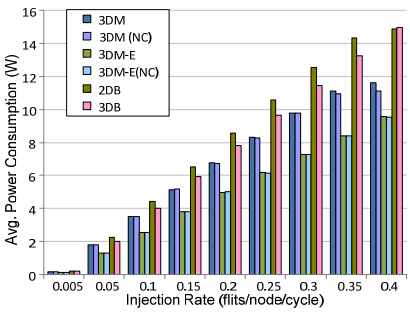
[Memory Configuration]

Private L1 Cache:	Split I and D cache, each 32 KB 4 way set associative, 64 B line, 3-cycle access time
Shared L2 Cache:	Unified 14 MB with 28 512 KB banks, each bank 4 cycle access (assuming 2GHz clock)
Memory:	4 GB DRAM, 400 cycle access, Each processor up to 16 outstanding memory requests

MIRA: Performance Analysis



(a) Uniform Random (UR)



(b) Uniform Random (UR)

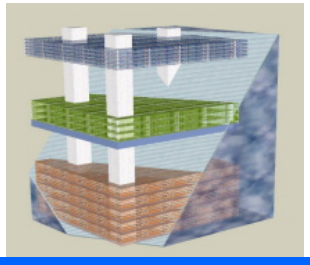
Latency

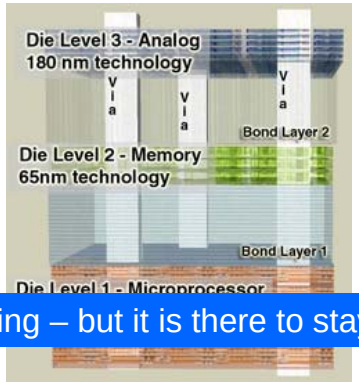
- 3DM-E is 26%, 51%, and 49% better than 3DB, 2DB, and 3DM, respectively in UR (inj. rate:0.3)
- 3DM/3DM-E performs better than 2DB/3DB in both NUCA and MP-Trace
- Pipeline Combination in 3DM/3DM-E helps improving performance

Power:

- 3DM router has lower power consumption than 2DB (22%) and 3DB (15%)
- 3DM-E has lower power consumption than 2DB (42%) and 3DB (37%)

Conclusion





There will be more forms of Stacking – but it is there to stay