# Case Study - NUMAchine

Zeljko Zilic
McConnell Engineering Building
Room 536


## Outline

- NUMAchine objectives and background
- Hardware organization
- Cache coherence
- Implementation details
- Performance summary and analysis

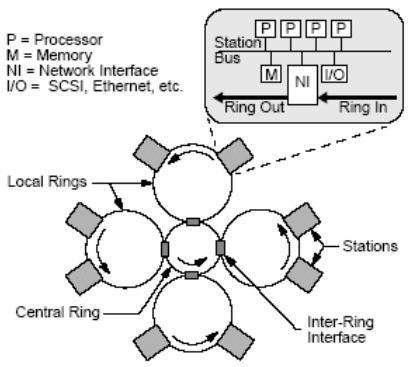
# NUMAchine Objectives

- Unique University research project
- Scalable, yet simple (feasible) & efficient cache coherent machine design
  - Directory-based CC, topology-specific
  - Programmable logic use
- OS Innovation: Tornado -> IBM K42
- Parallel compiler research



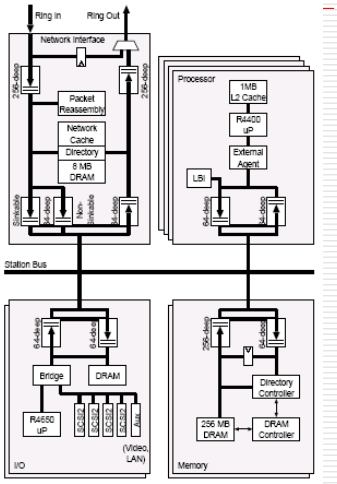
# Organization Overview & History

- Hierarchical ring
  - Research (theoretical) on rings with short packets - Hamacher, Vranesic & Zaky, also K. Sevcik
  - LAN: Tornet
  - Tornet II (A. Sanwalka)
  - Hector (R. White, D. Lewis)
    - K. Farkas Ph. D.
- Clustered: bus at lowest level of hierarchy
- Station



# High-Level HW Organization

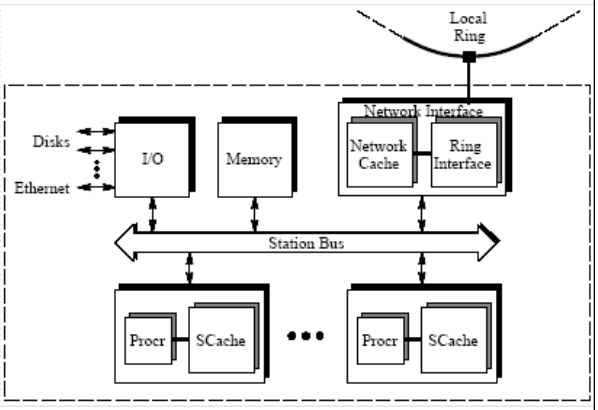
- Implemented using programmable logic
  - CPLDs and FPGAs
  - Flexible
  - Quick prototyping
  - Inexpensive
  - Hence: feasible
    - Stanford Flash machine never completed






# NUMAchine Station

- Bus
- No snooping
- 4 processors
- Memory
- NI
- IO Module



# Bus Organization

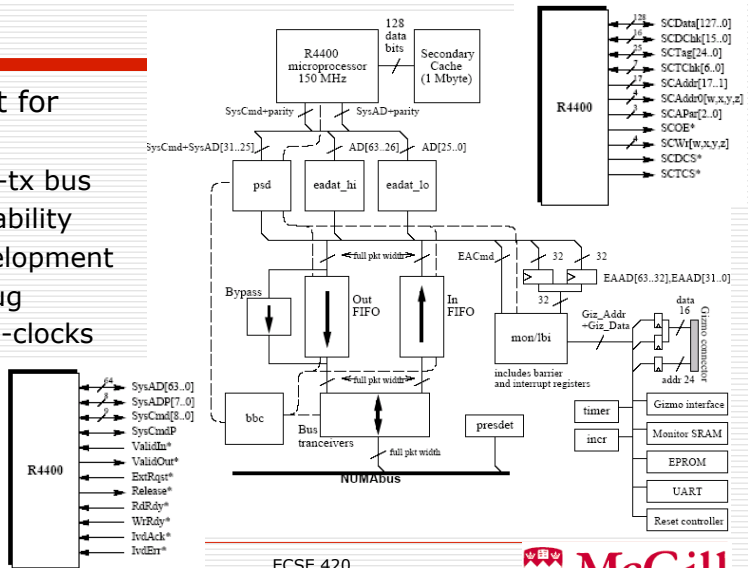
- Split-transaction
- Pipelined
- Electronics – Futurebus
- Bipolar BTL
  - Open collector
- Support for CC protocol
  - (Non)Sinkable
  - Backpressure

| | | | | | |
|---|---|---|---|---|---|
| packet | positive | Sender ID | | | 4 |
| | | Command + parity | | | 18 |
| | | Response Select | | | 9 |
| | | Address/Data | parity | | 8 |
| | | | Data | 64 | address 40 |
| | | | | | destination station mask 8 |
| | | | | | monitoring phase ID 4 |
| | | | | | source processor/device ID 4 |
| | | | | | source station mask 8 |
| arbitration | positive | Bus Request | | | 1 |
| | | Bus Hold[†] | | | 1 |
| | | Bus Grant | | | 1 |
| | | Bus Busy | | | 1 |
| | | Busy Sinkable | | | 9 |
| | | Busy Nonsinkable | | | 9 |
| | | Select | | | 9 |
| miscellaneous | positive | Station/Ring ID | | | 4 |
| | OC (neg) | Bus Reset | | | 1 |
| | OC (neg) | Bus Reset Request | | | 1 |
| | OC (neg) | System Reset Request | | | 1 |
| | positive | Cache Line Size | | | 1 |
| | | Memory Present | | | 2 |
| | | IO Present | | | 2 |
| | PECL | Bus Clock | | | 2 |
| | positive | Uart Data | | | 1 |
| | | Uart Poll Request | | | 1 |
| | | GA Test | | | 4 |
| monitoring | positive | Local Ring Busy | | | 1 |
| | | Global Ring Busy | | | 1 |



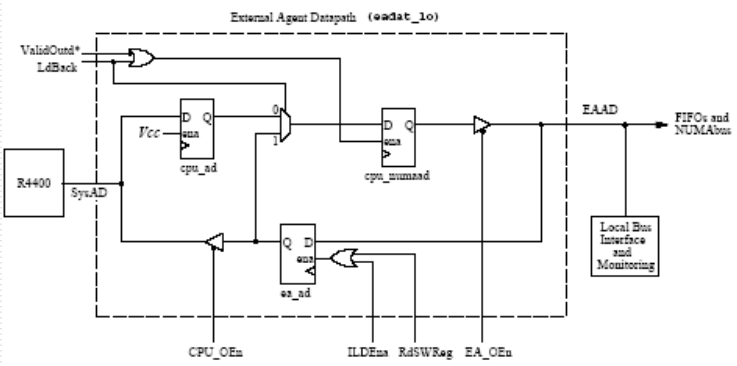
# NUMAchine Processor Board

- Support for
  - CC
  - Split-tx bus
  - Scalability
  - Development
  - Debug
  - Multi-clocks
- MIPS R4400



# External Agent

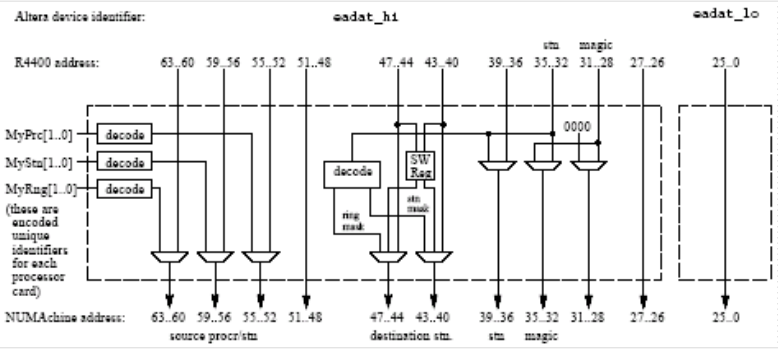
- MIPS Interface + Debug/monitoring



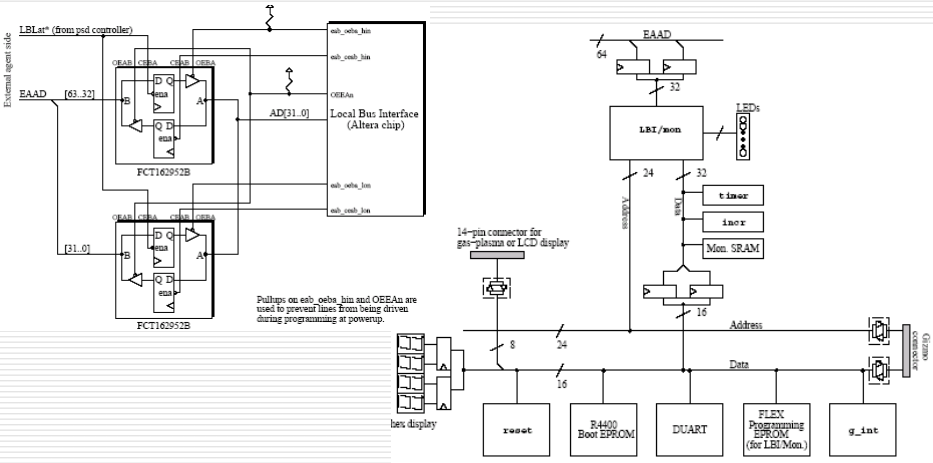



# Data Path Manipulation

- Embedding into NUMAchine protocol



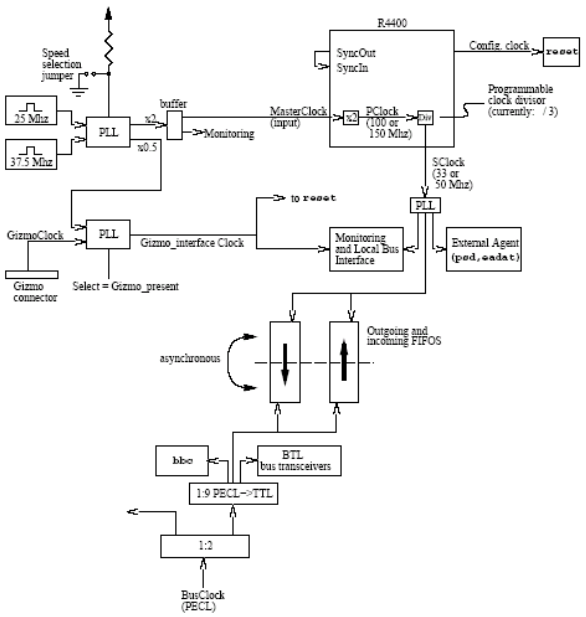
# Local Bus Interface





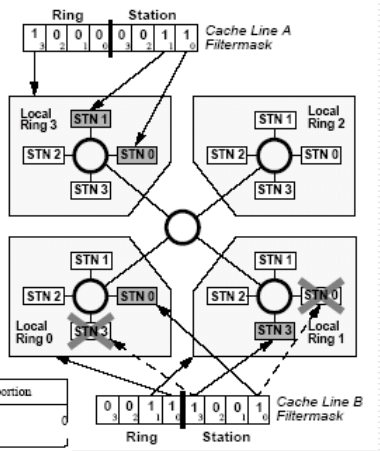

# Clocks

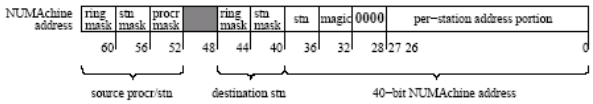
- Multiple-clock domains
- Synchronized to the rest
  - Synchronization between clock domains



# Naming, Addressing, Encoding

- Filtermasks – unique address
  - Multicast
  - Reduce unnecessary invalidations

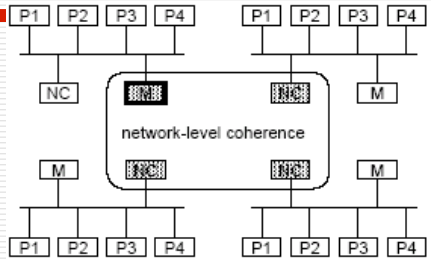


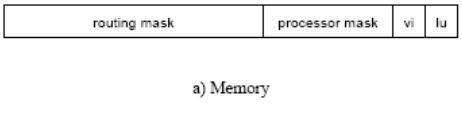



# Cache Coherence Overview

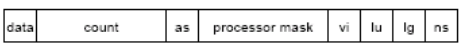
- Writeback, invalidate
- Directory
  - Hierarchical
  - Memories, Network Cache
- Cache Line states




| | |
|---|---|
| *Local Valid (LV)* | One or more processors caches within the station have a shared copy. Remote stations do not have valid copies. |
| *Local Invalid (LI)* | One local processor cache has a modified copy. There are no other valid copies. |
| *Global Valid (GV)* | One or more remote stations have a shared copy, and there may also be local copies in processor caches within the station. |
| *Global Invalid (GI)* | One remote station has a modified copy, and there are no other valid copies. |



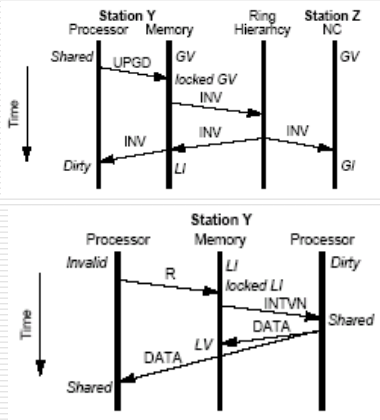

a) Memory




b) Network cache

# CC Protocol – Local Operations

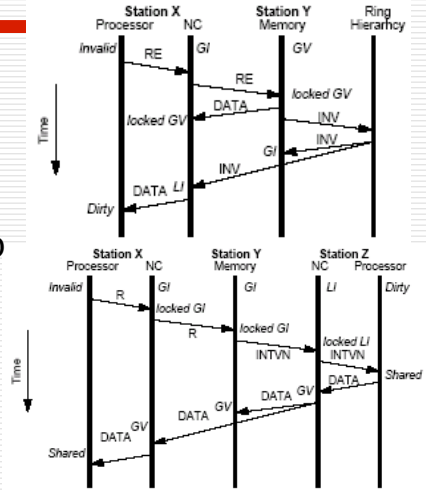
- Local write operation (home memory on the same station)
  - Upgrade upon GV
    - Converted to Inv, locked
    - Serialized on root ( "height" inferred from directory data)
- Local read
  - Invalidation if LI
    - *Intervention* – gets data from a processor indicated by directory
  - Uses locking to shorten invalidation cycle






# CC – Remote Ops

- Remote write
  - Needs to go to NC
  - Inv upon reaching home location
  - Serialized through top (root) of hierarchy
- Remote read
  - Needs second NC – remote
  - Remote intervention

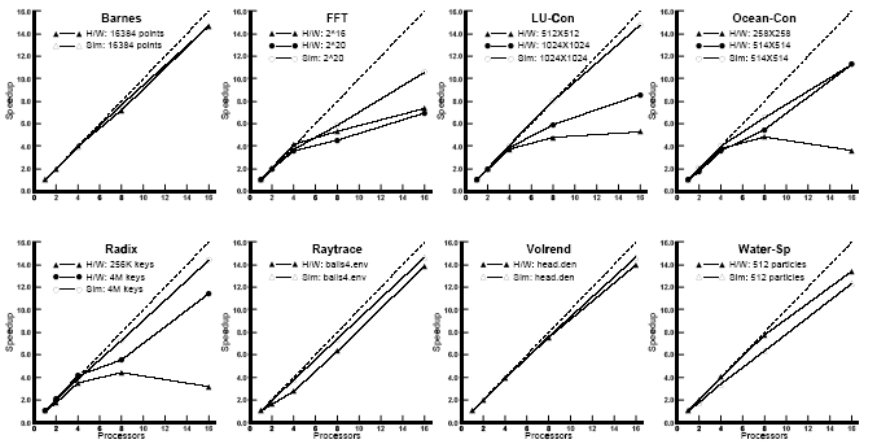


# More on NUMAchine CC

- Global ordering (sequencing) of invalidations -> sequential consistency
  - All processors see writes in the same order
- Don't need to wait for global ack on Inv.
  - Enough receive 1 inv, ordering takes care of the rest
- More information on cache coherence:
  - A. Grbic, Ph. D. thesis
  - S. Srbljic, report on profiling memory accesses
- More on design:
  - R. Grindley, Ph. D. thesis
  - K. Loveless, M. Eng. Thesis
  - M. Gusat, M. Eng. thesis
  - NUMAchine web page eecg.toronto.edu/parallel/parallel/numadocs.html



# Speedups Observed



# Interconnect Utilization

- SPLASH-2 benchmark suite
- Interconnect
  - Bandwidth suitable
  - Latency tolerable
  - Bisection bandwidth secondary concern

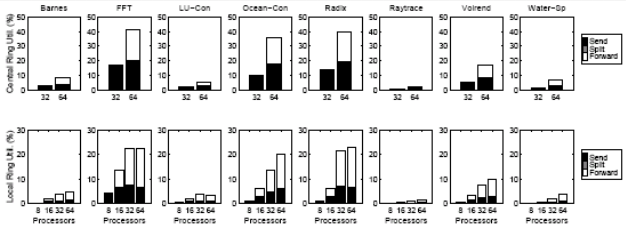



Figure 4. Central and Local Ring utilizations

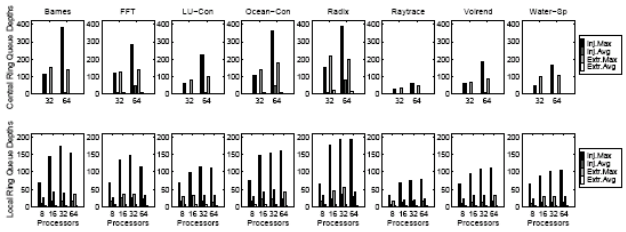
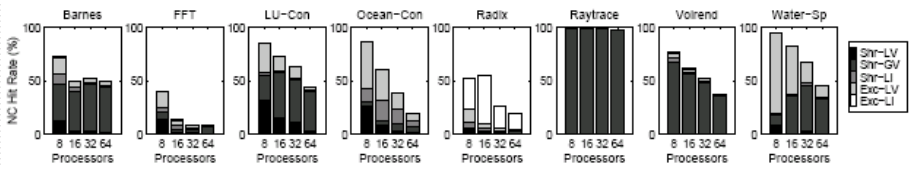



Figure 5. Central and Local Ring queue depths



# Miss Rates and Latency

- Network Cache Miss Rates




- Latency, app summary

| level of hierarchy | 150-MHz PCLKs | 50-MHz SCLKs |
|---|---|---|
| L1 cache | 1 | *n/a* |
| L2 cache | 6 | *n/a* |
| Local memory | 135 | 45 |
| Local network cache | 165 | 55 |
| Other L2 cache | 255 | 85 |
| Rem. mem. (same ring) | 594 | 198 |

| Benchmark Characteristics | | | Observations | |
|---|---|---|---|---|
| name | c-to-c ratio | locality | ring util. | speedup |
| Barnes | low | good | low | good |
| FFT | high | poor | higher | poor |
| LU-Con | low | good | low | poor |
| Ocean-Con | high | moderate | higher | good[1] |
| Radix | high | poor | higher | good[1] |
| Raytrace | low | moderate | low | good |
| Volrend | low | moderate | medium | good |
| Water-Sp | low | good | low | good |

# Coherence Impact

- Measured impact of coherence protocol on performance

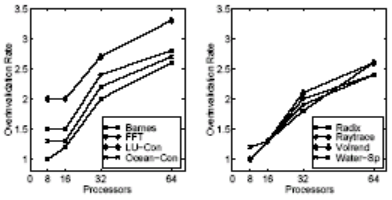



Figure 6. Overinvalidation rates

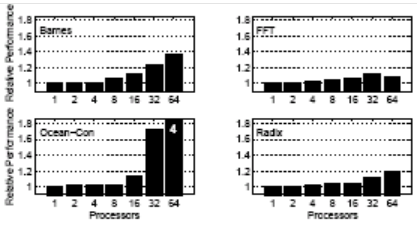



Figure 7. Turning off cache coherence



# Was it Worth it?

- Last (or so) large computer system designed at Universities
  - Hardware
  - OS (borrowed some from SGI Irix)
  - Compilers
  - Apps: Chess machine rated at ~2300
- Feasible, but only with programmable logic implementation
- Excellent opportunity for graduates, industry employing them