PicoServer : Using 3D Stacking Technology To Enable A Compact Energy Efficient Chip Multiprocessor

Taeho Kgil, Shaun D'Souza, Ali Saidi, Nathan Binkert, Ronald Dreslinski, Steve Reinhardt, Krisztian Flautner, Trevor Mudge

Advanced Computer Architecture Lab, University of Michigan,
In collaboration with HP Labs, Reservoir Labs, ARM

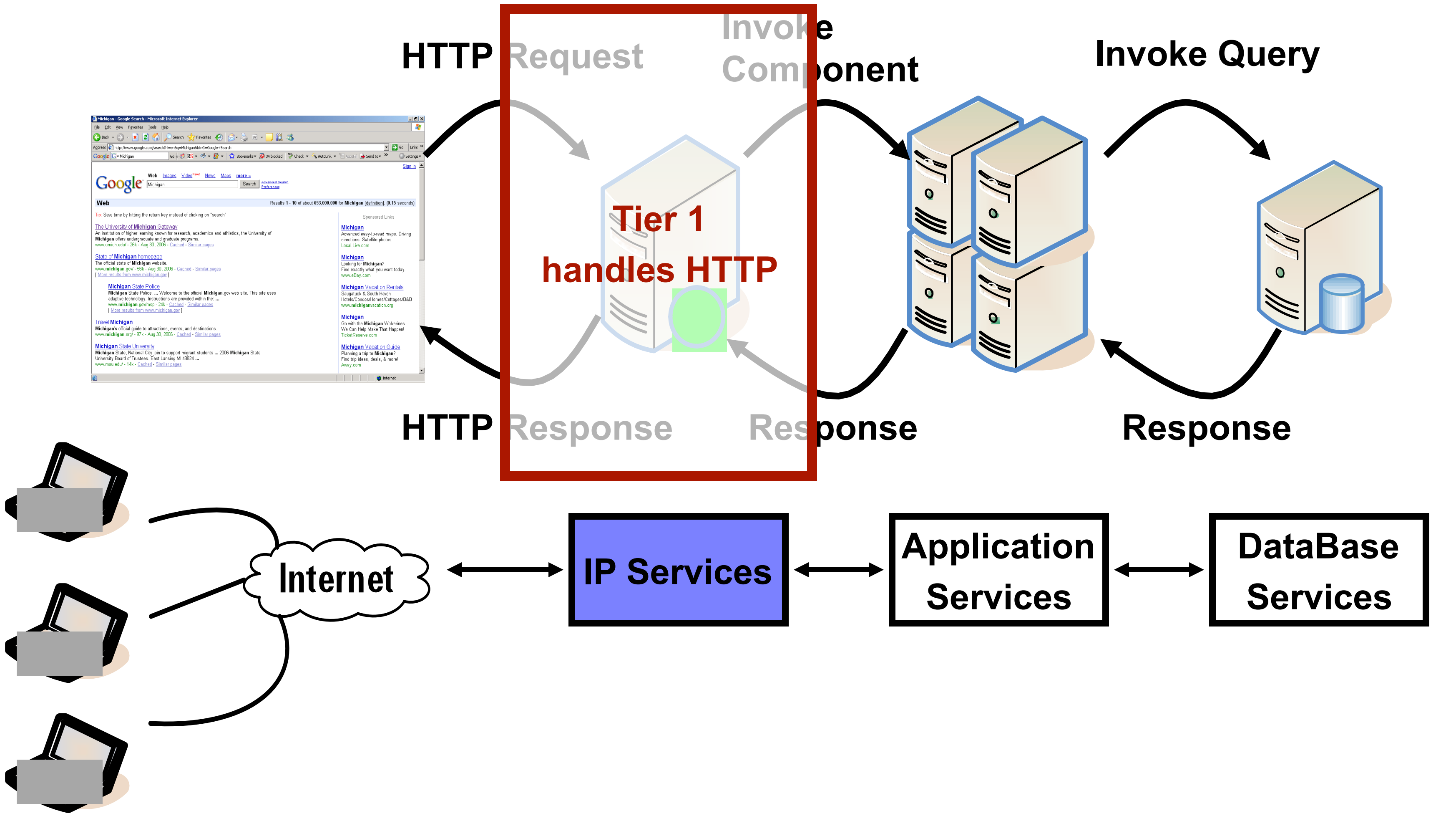
Motivation

- Vast amounts of servers required
  - AOL, Google, Yahoo maintain large datacenters
  - General purpose processors not efficient to handle server workloads
- Opportunities with 3D stacking technology
  - Extreme integration
  - Improved throughput and latency
- Leverage 3D IC to build energy efficient Tier 1 servers
  - Tier 1 workloads require high memory throughput and modest ILP
  - CPU, Memory Controller, NIC, on-chip DRAM altogether in a single package

Outline

- **Background**
  - **Server platform**
  - **Advances in technology - 3D stacking & DRAM**
- PicoServer Architecture
- Methodology
- Results
- Conclusions and Future Work

3 Tier Architecture

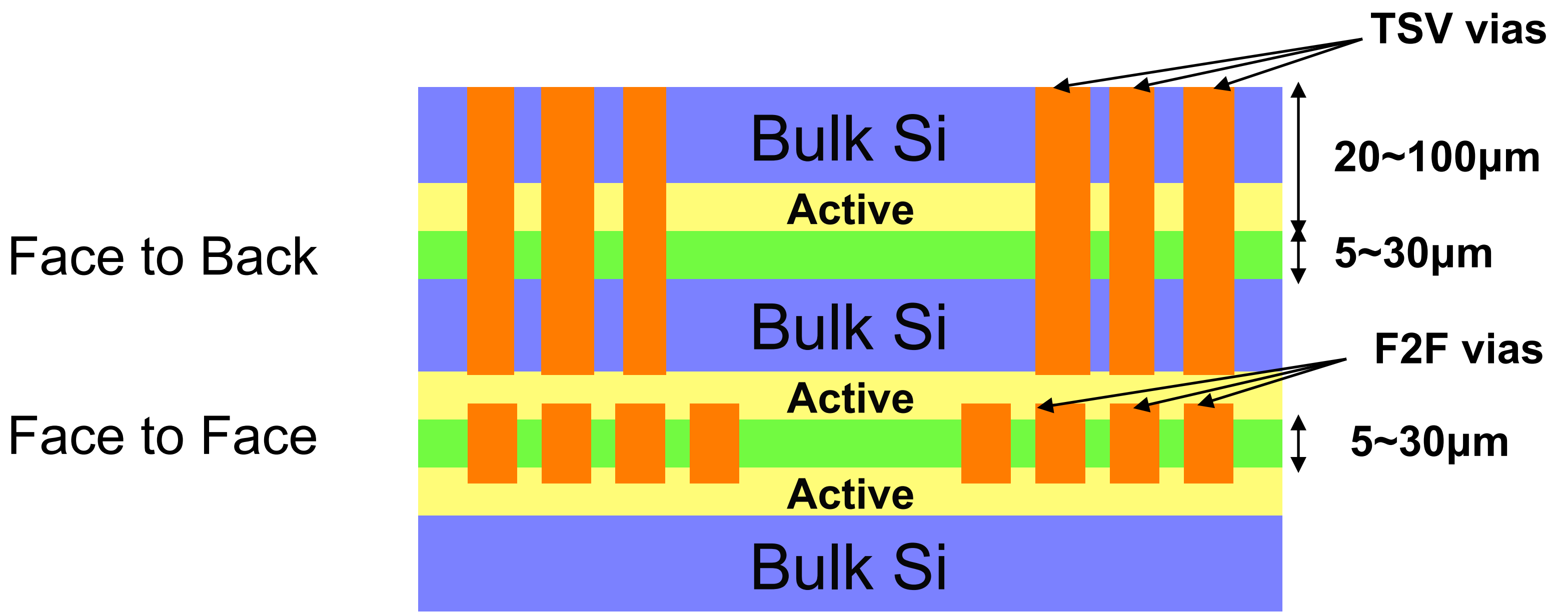


Behavior of Commercial Server Workloads

Attribute	Web99	SAP 2T	TPC-H	TPC-C
Application Category	Web Server	ERP	DSS	OLTP
ILP	Low	Med	High	Low
TLP	High	High	High	High
Working-set size	Large	Med	Large	Large
Data-sharing	Low	Med	Med	High

From S.R Kunkel et al, IBM J. R&D vol. 44 no.6, 2000

What is 3D stacking technology? – using 3D vias to connect multiple dies



3D stacking pros and cons

- High bandwidth (throughput)
  - Millions of die to die connections
- Reduces interconnect length
  - Interconnect becoming a problem as feature sizes shrink
- Extreme integration of components manufactured from different process technology
  - DRAM, Flash Memory, Analog, RF circuits etc
- Thermal problems
  - Power density limits the number of stacks
- Chip verification & Yield
  - Verification at the die, wafer and post-package level is necessary
  - Overall Yield is a product of individual die yield and 3D stacking yield

Roadmap for 3D stacking and DRAM - Where are we?

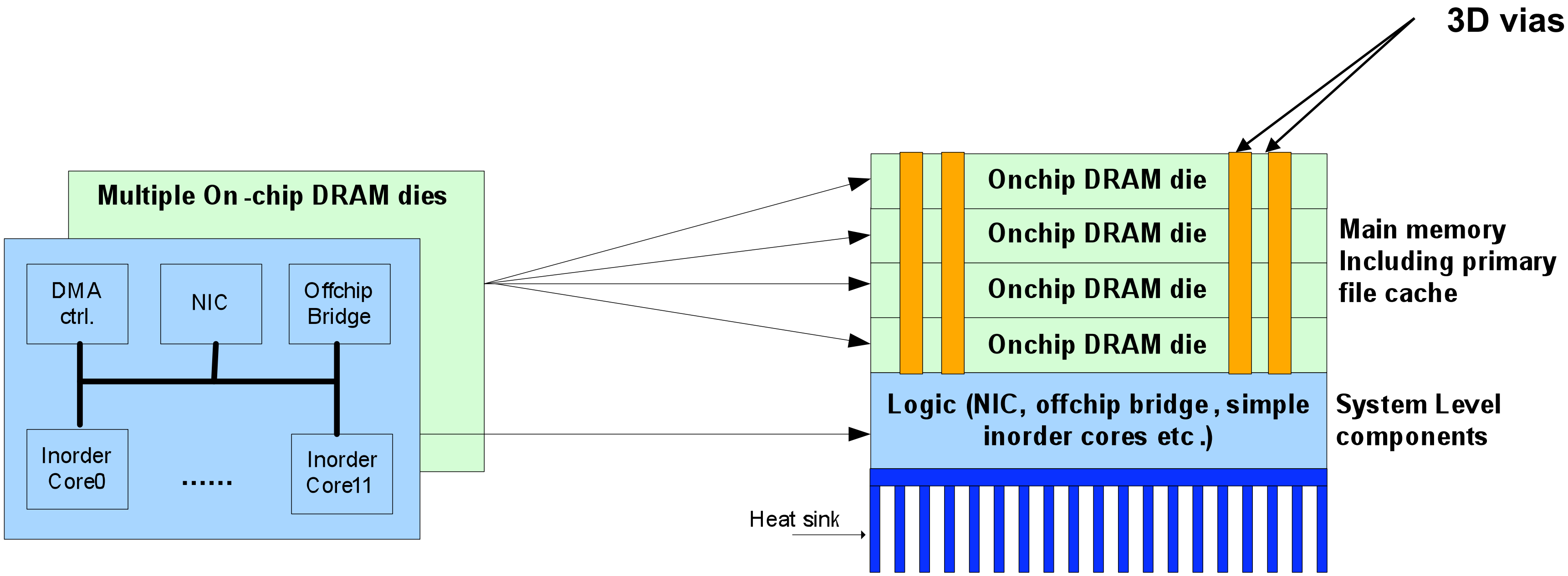
	2005	2007	2009	2011	2013
Number of stack max. for low-cost / handheld - 3W power budget	6	7	9	11	13
Number of stack max. for high performance	2	3	3	4	5
Cell Density of SRAM MBytes / cm^2	11	17	28	46	74
Cell Density of DRAM MBytes / cm^2	153	243	458	728	1,154

From ITRS 2005 Roadmap

Outline

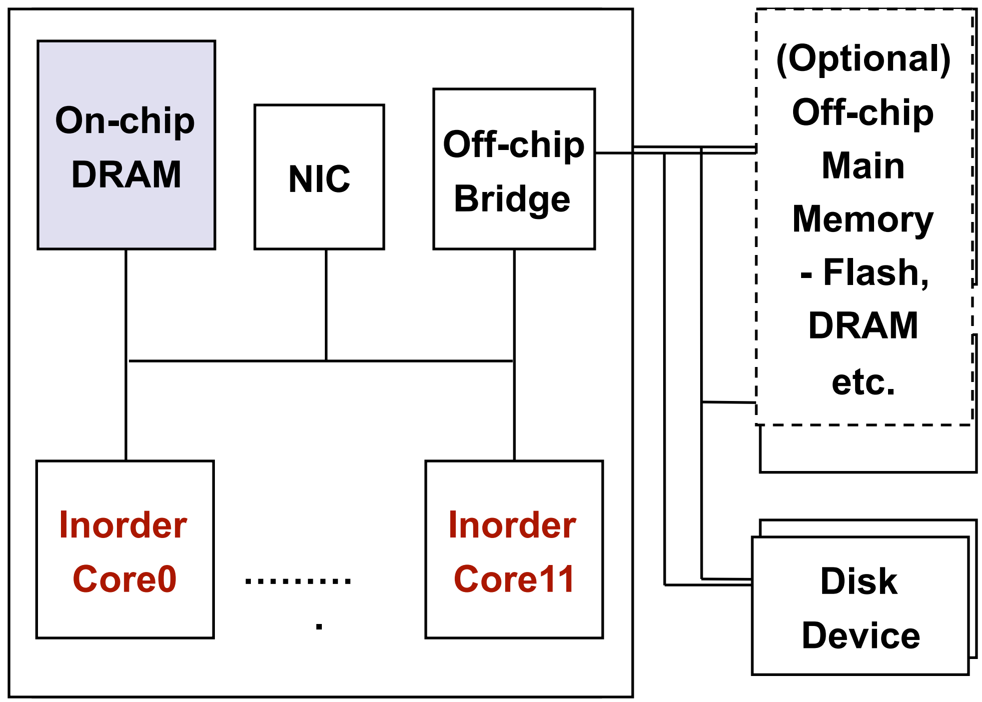
- Background
- **PicoServer Architecture**
  - **Overall Architecture**
  - **Architecture of logic components**
  - **Architecture of interconnect**
  - **Role of on-chip memory**
- Methodology
- Results
- Conclusions and Future Work

PicoServer Architecture – Using simple cores with simple interconnect



Logic to Memory – F2F via, Memory to Memory – TSV via

Extreme integration and NUMA

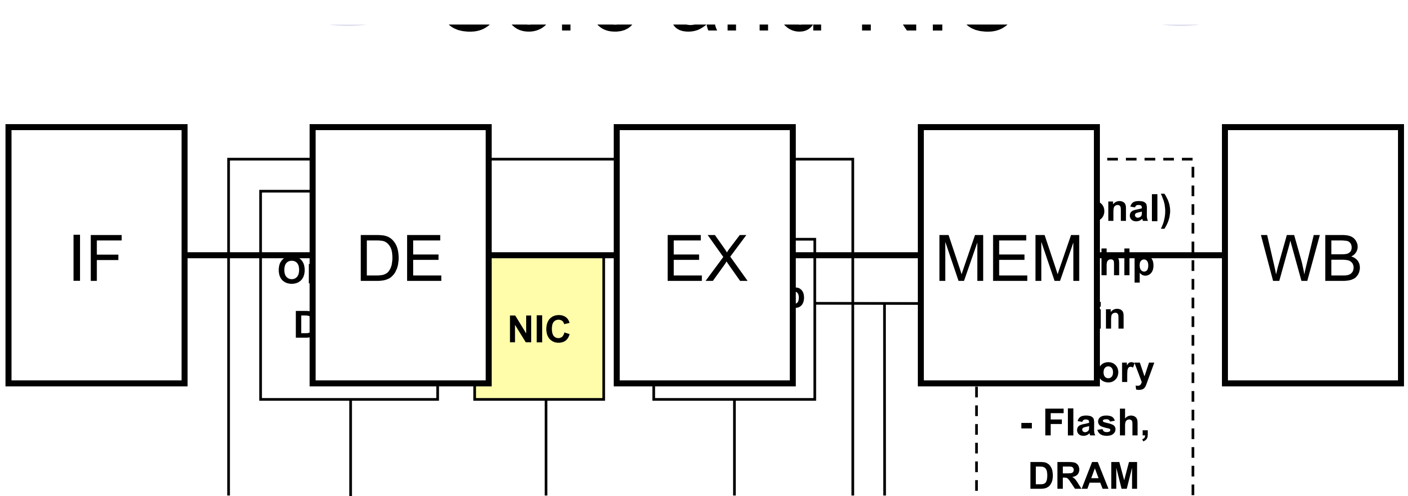


CMP with 3D stacking

PicoServer and 3D stacking

- No need for L2 cache
  - Access latency and bandwidth of on-chip DRAM similar to a L2 cache
  - Additional cores can replace the L2 cache
- High performance low power interconnect
  - High bandwidth memory to core interface
  - The added degree of freedom reduces interconnect length
- Multicores clocked at modest frequency (500MHz)
  - Tier 1 server workloads are not computationally intensive
  - TLP more of an issue
- On-chip memory
  - Server applications → on-chip DRAM
  - Hundreds of MB of DRAM can be integrated on-chip
    - Additional memory can be available externally

Using Scalar Cores and Intelligent NICs

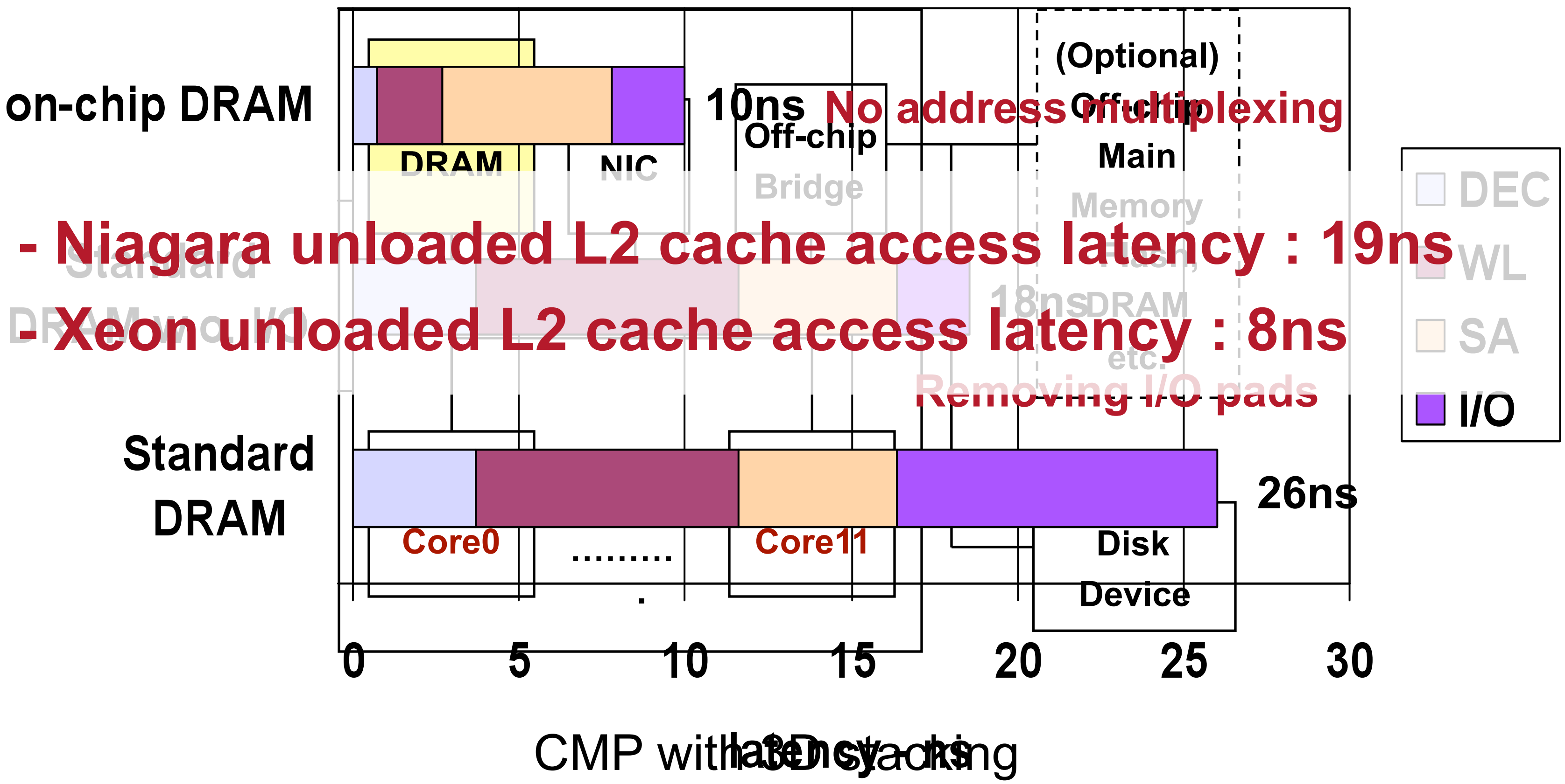


- Simple 5 stage pipeline clocked at low frequency – 500MHz
  - Maintain a reasonable power density to stack many die layers.
  - Opportunities to use low power process technology and DVS
- Standard branch predictor
  - 90 ~ 95% branch prediction
- ISA support for multicores
- Integrated DRAM controller per core to interface with on-chip memory
- Intelligent NICs are required to do load balancing
  - Load balancing achieved with Microsoft RSS like methods

Shared simple interconnect

- More than 70% of interconnect traffic is due to cache misses
  - Interconnect should handle **cache miss traffic** better than other types of traffic.
- Low frequency wide bus provide high throughput & low transfer latency
  - 3D stacking enables high throughput low frequency interconnect to on-chip DRAM
  - Simulations suggested a wide shared bus produced sufficient performance
    - Minimal queue delay in wide shared bus

The role of on-chip DRAM



- Niagara unloaded L2 cache access latency : 19ns
- Xeon unloaded L2 cache access latency : 8ns

R. Matick IBM

The role of on-chip DRAM (cont.)

- A large portion of main memory is used as disk cache
  - Less than 64MB occupied by application, OS
  - Similar memory usage also reported in many server applications
- 100’s of MB of on-chip DRAM is enough to hold code & data and a portion of disk cache

Outline

- Background
- PicoServer Architecture
- **Methodology**
- Results
- Conclusions and Future Work

Methodology

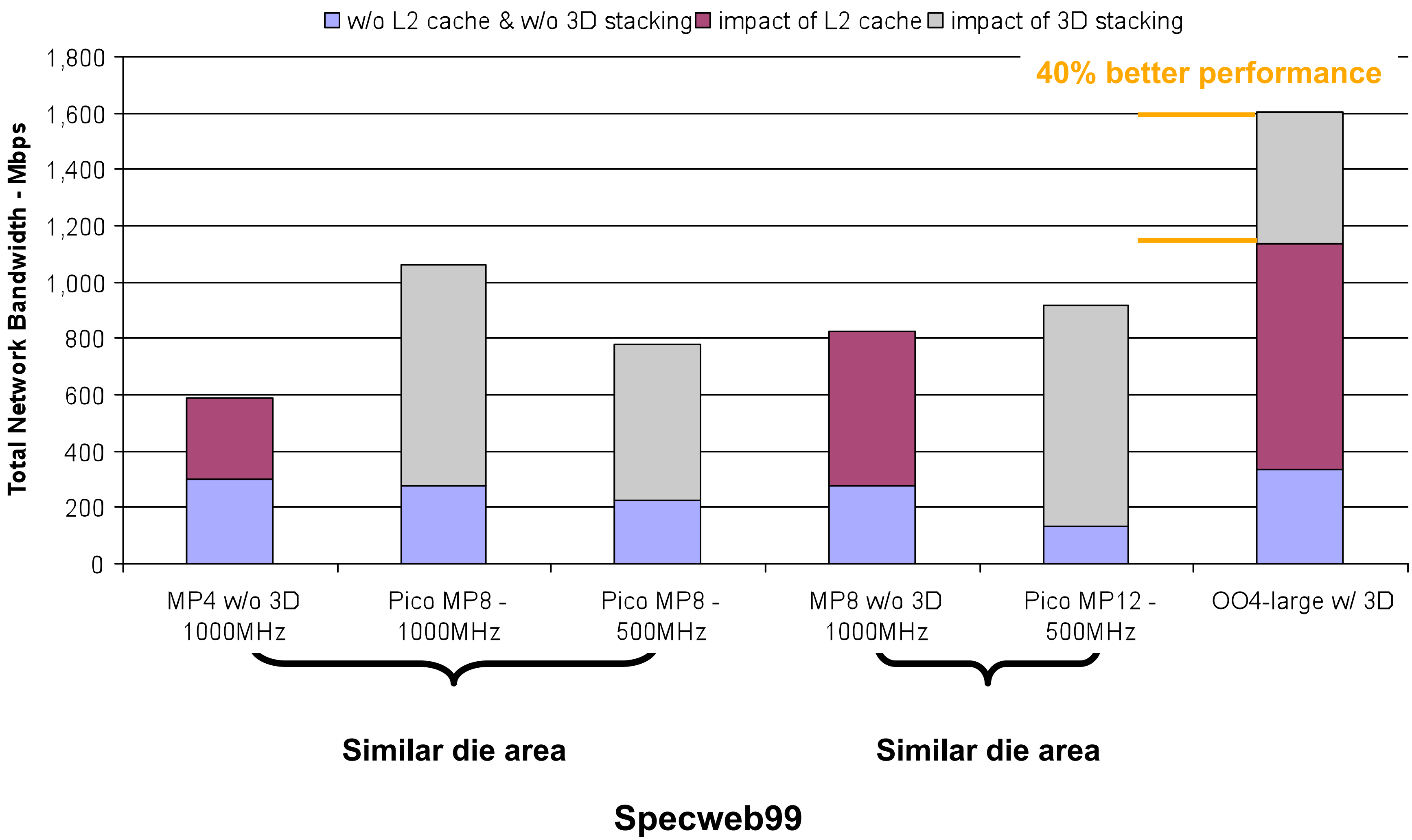
- Full-system simulator M5
  - Models client-server connection
  - Generated client requests that saturate processor utilization in the server
- SURGE (static web), SpecWeb99 (dynamic web), Fenice (video streaming) and dbench (file serving) for Tier 1 server workloads
- Relied on empirical measurements from ISSCC, IEDM papers and datasheets to estimate power
- Calibrate empirical measurements with ITRS roadmap predictions, scaling rules and analytical FO4 model (for processor)
  - Overestimate most values to be on the safe side

Outline

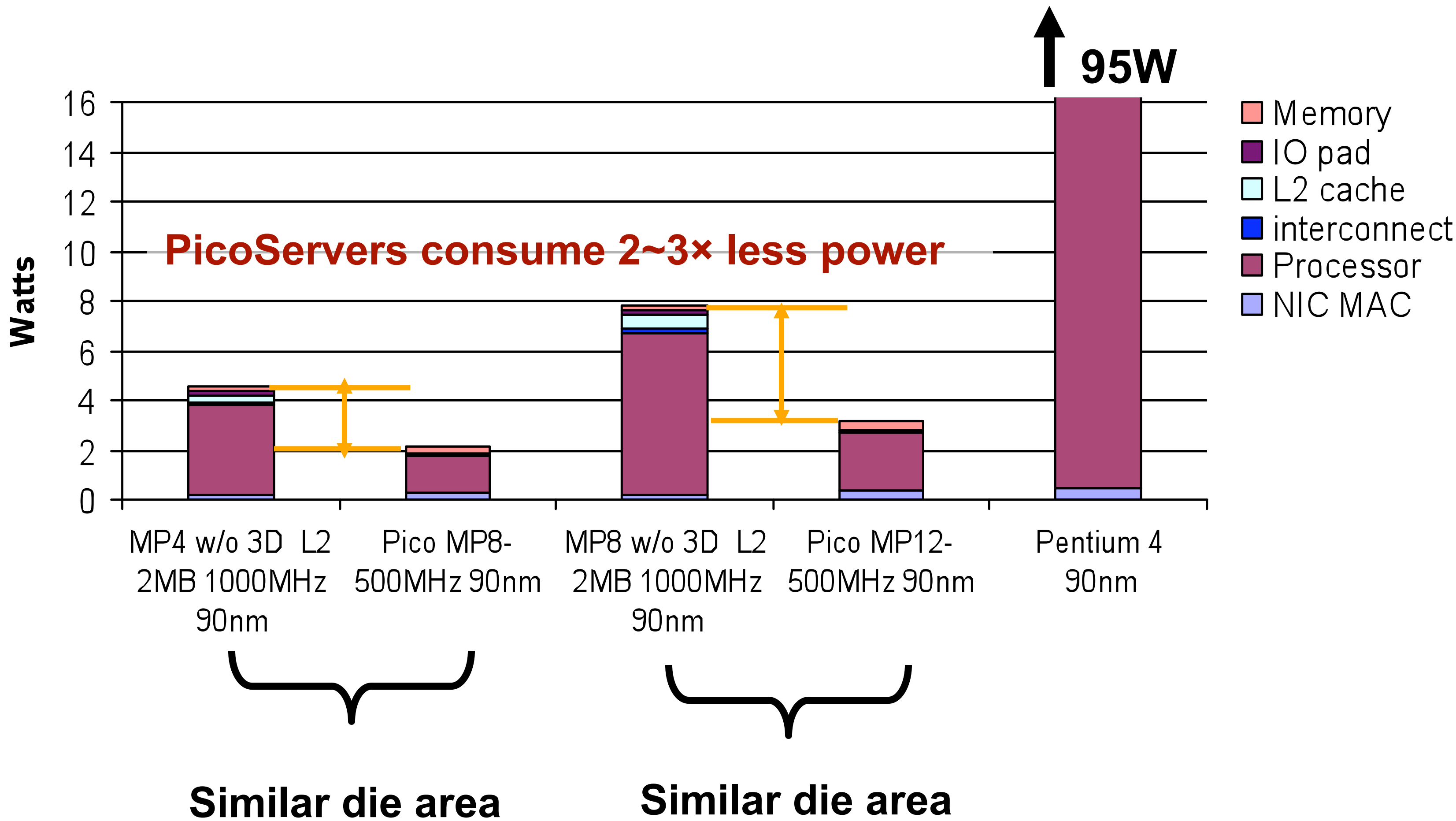
- Background
- PicoServer Architecture
- Methodology
- **Results**
  - **Overall Network Bandwidth - Mbps**
  - **Overall Estimated Total Power**
  - **Energy Efficiency**
- Conclusions and Future Work

Additional cores yield improvement in Network Performance while operating at half the frequency

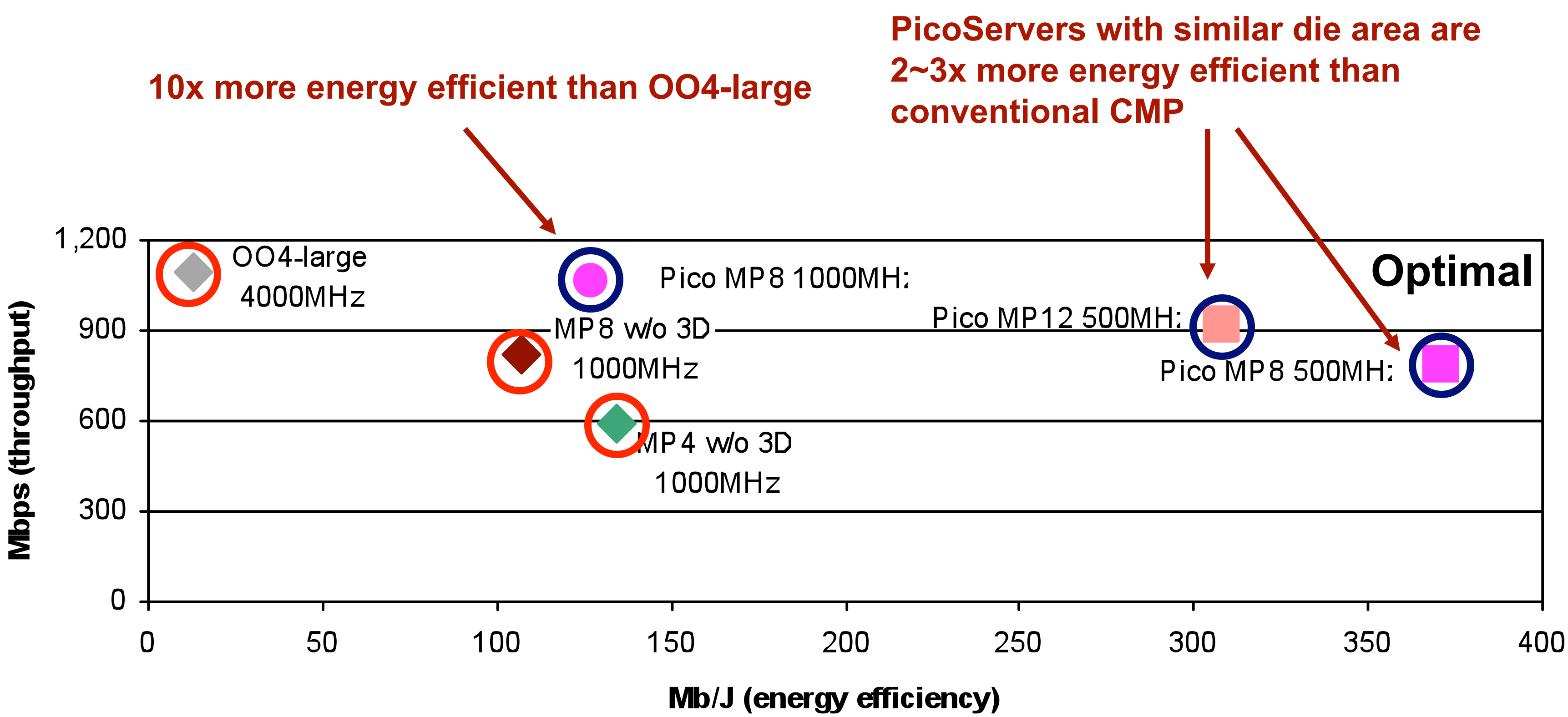
Overall Network Bandwidth – Mbps



Overall Estimated Total Power



Energy Efficiency Pareto Chart



Specweb99

Outline

- Background
- PicoServer Architecture
- Methodology
- Results
- **Conclusions and Future Work**

Conclusions & Future Work

- 3D stacking complements Tier 1 server workloads
  - High throughput memory bandwidth
  - More Processing Elements on die
  - Extreme integration for small form factors
- Simple multicores generate acceptable network bandwidth while consuming low power
  - For a 3W budget, 0.6~1.4Gbps network bandwidth
- Future Work
  - Investigate core architecture for computation intensive server workloads
  - Investigate energy efficient NUMA architectures for datacenter platforms

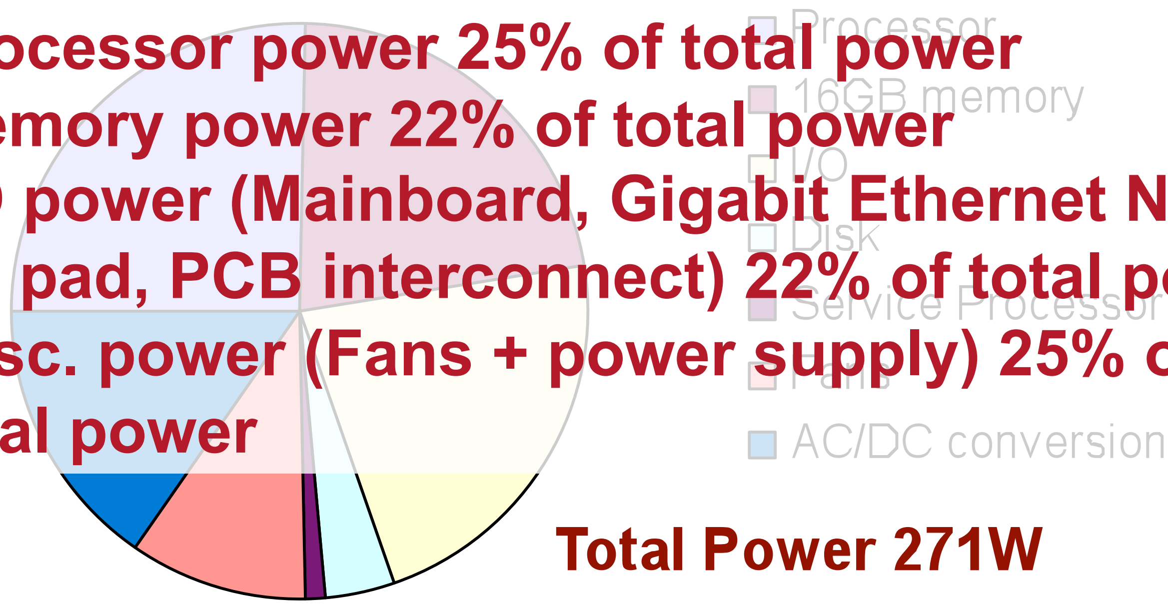
Questions???

- ?????

Backup Slides

System Level Power consumption

SunFire T2000 Power running SpecJBB



Power-wise

- **Processor power 25% of total power**
- **Memory power 22% of total power**
- **I/O power (Mainboard, Gigabit Ethernet NICs, I/O pad, PCB interconnect) 22% of total power**
- **Misc. power (Fans + power supply) 25% of total power**

Total Power 271W

From Sun talk given by James Laudon

3D via parameters

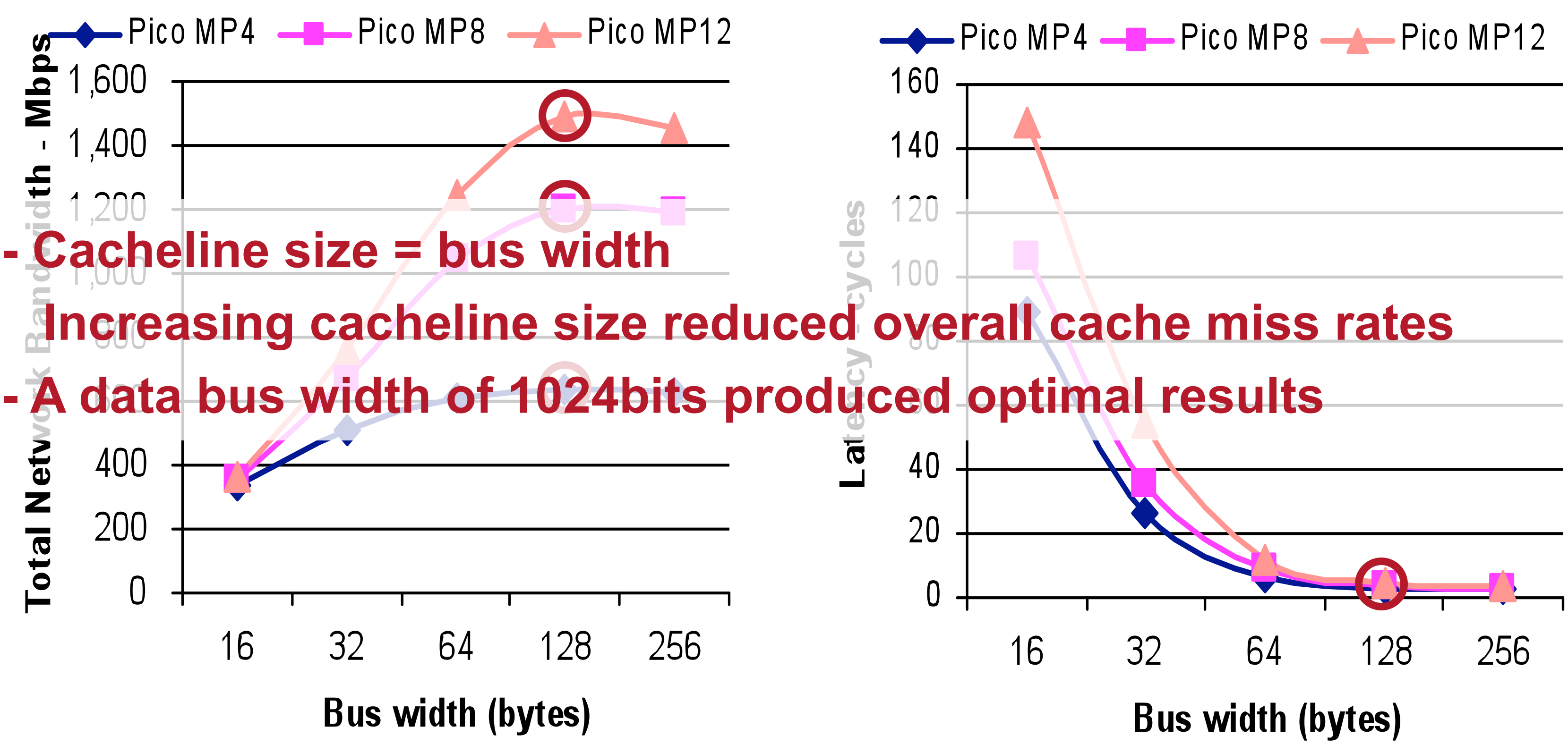
	Tezzaron 2nd Generation	Tezzron Face to Face	RPI	MIT 3D
Size	1.2μ x 1.2μ	1.7μ x 1.7μ	2μ x 2μ	1μ x 1μ
Minimum Pitch	[illegible]	[illegible]	[illegible]	N / A
Feed Through Capacitance	2~3fF	<<	<<	2.7fF
Series Resistance	<0.35Ω	<	<	<

A 3D via delivers minimal delay overhead & about the size of a 90nm 6T SRAM cell.

Via density exceeds 14,000/mm^2

Numbers from Tezzaron Semiconductor, RPI, MIT

Evaluation of a Wide shared Bus



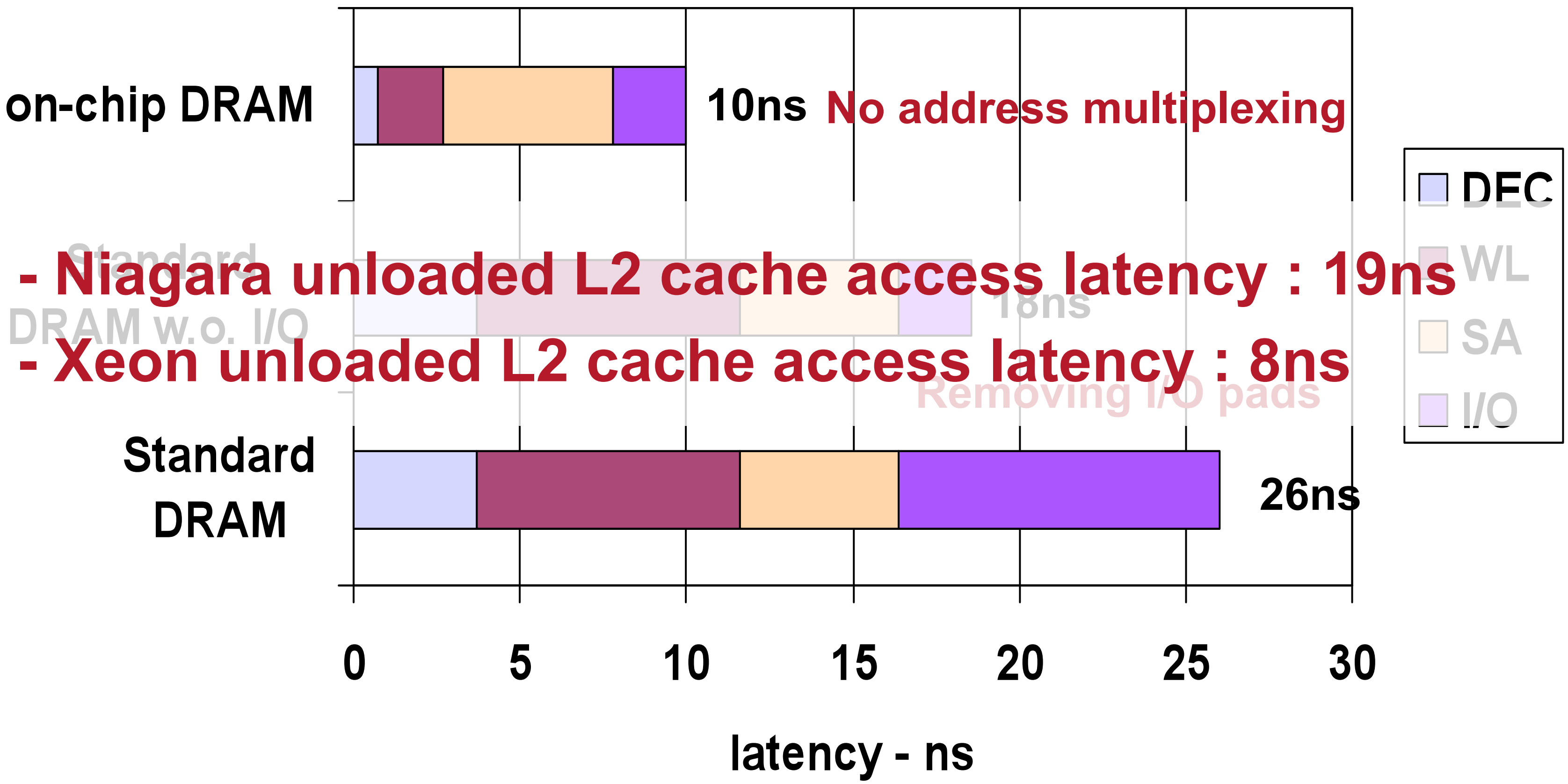
- Cacheline size = bus width

Increasing cacheline size reduced overall cache miss rates

- A data bus width of 1024bits produced optimal results

SURGE

The role of on-chip DRAM



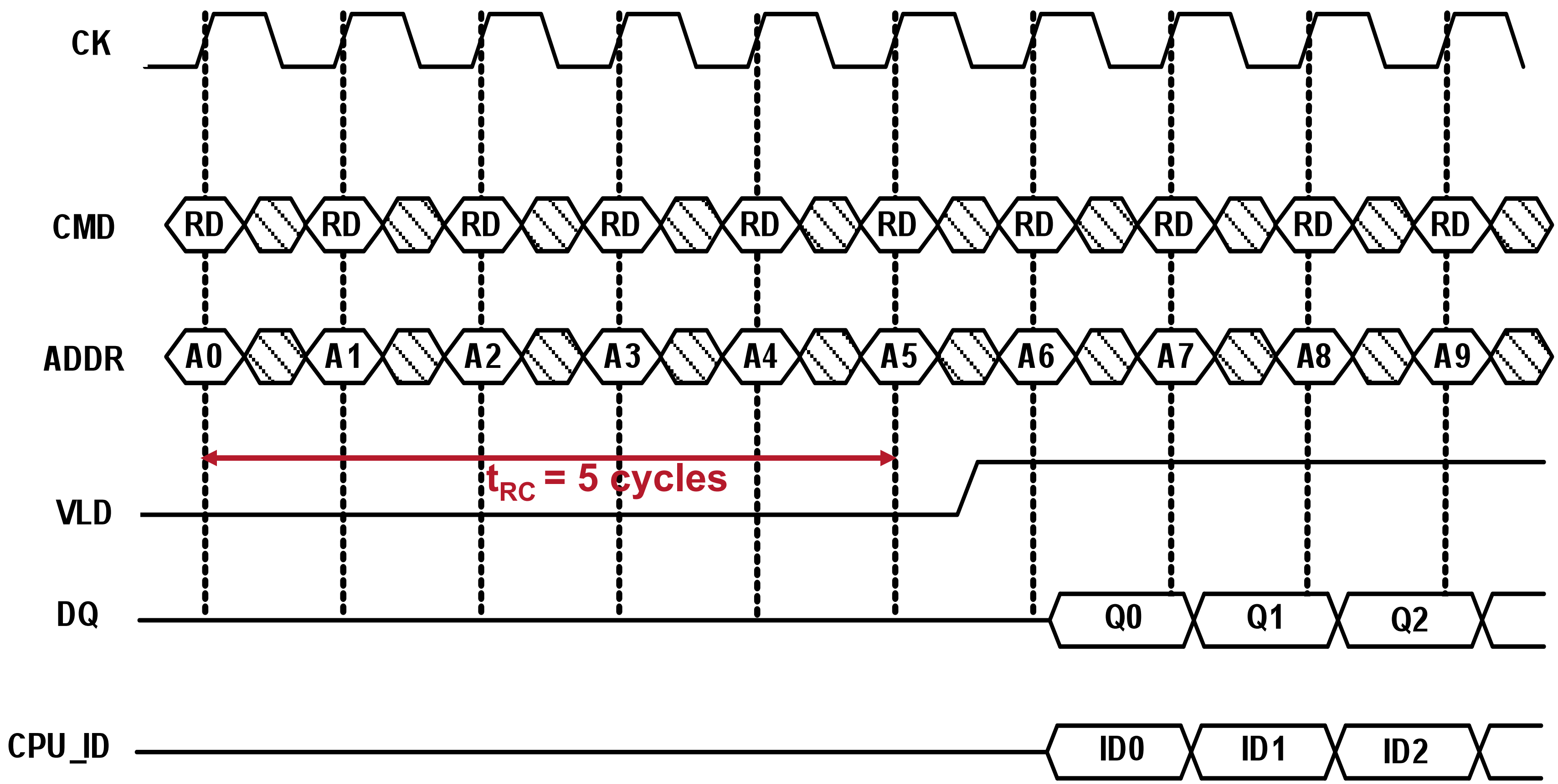
- Niagara unloaded L2 cache access latency : 19ns
- Xeon unloaded L2 cache access latency : 8ns

R. Matick IBM

Improving word line delay

- Word line delay depends on the resulting RC caused by the large number of gates
- One solution in reducing RC delay is by dividing the word line into smaller sections and to add buffers.
  - However, additional drivers and buffers add area.
- Another solution is to route the word lines in metal rather than polysilicon or silicide.
  - Independent studies show that aluminum word lines reduce wordline delay by 3x [Tanabe92]
- On-chip DRAM enables one to reallocate die area that was previously assigned to I/O & address multiplexing to improving word line delay with the above solutions.

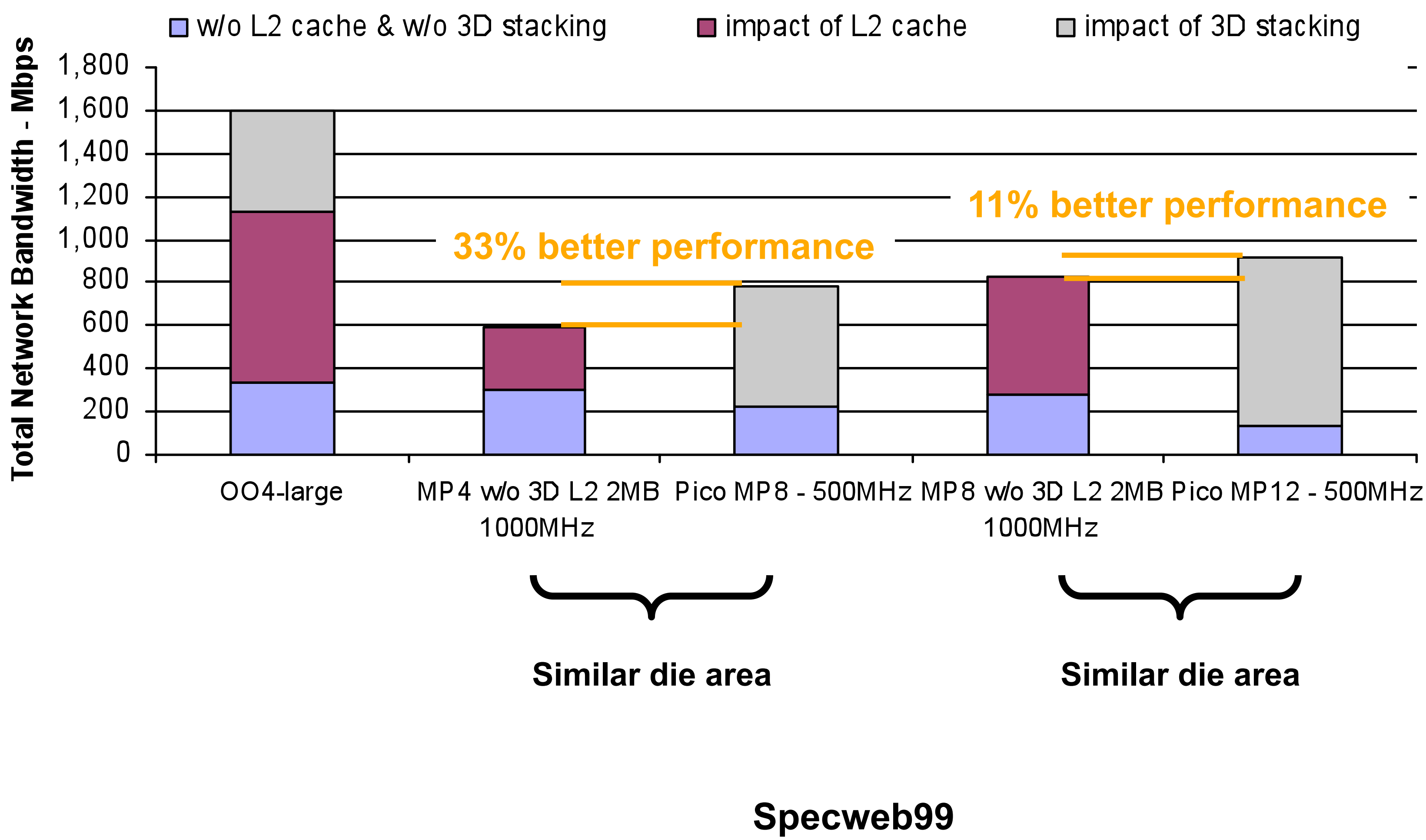
Example timing diagram – DRAM read



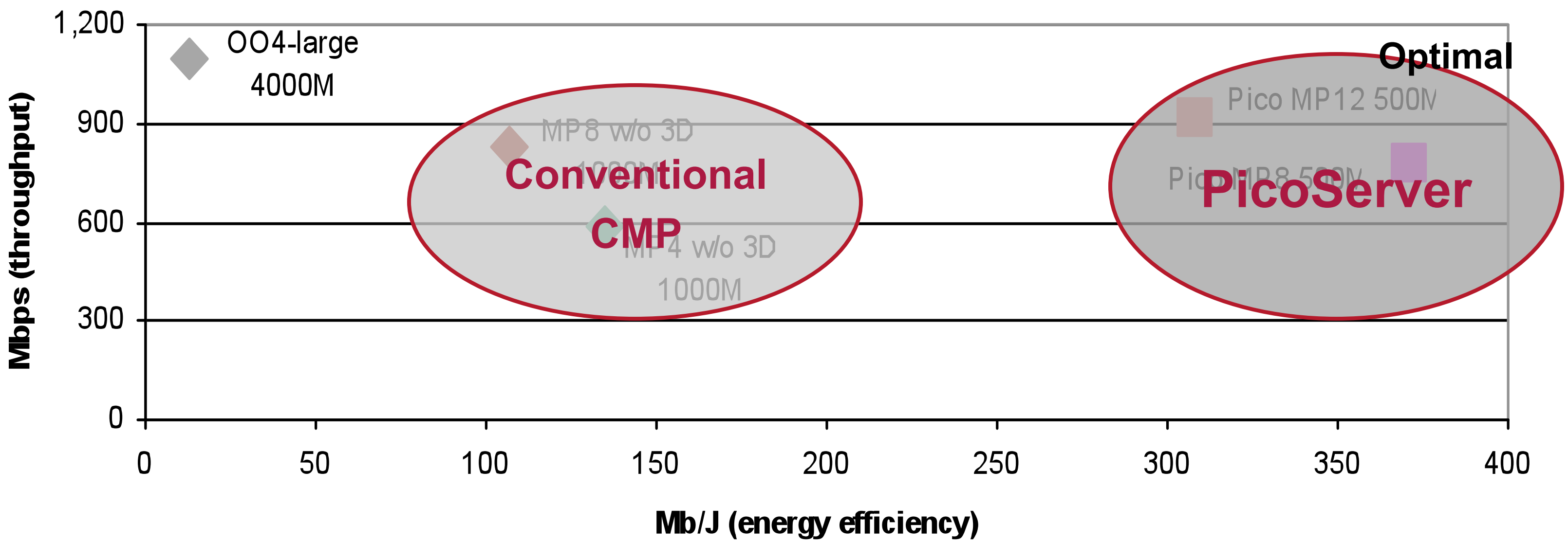
Commonly used configurations

	General Purpose Processor	PicoServer	Conventional CMP
Syntax	OO4-<small,large> w/ w/o 3D stacking	Pico MP<# of cores> – <freq>	MP <# of cores> w/o 3D stacking
Operating Frequency	4GHz	500MHz / 1GHz	1GHz
Number of Processors	1	4, 8, 12	4, 8
Processor Type	Out-of-Order	In-order	In-order
Issue width per core	4	1	1
L1 cache size	2 way 16KB or 128KB	4 way 16KB	4 way 16KB
L2 cache size	8 way 256KB or 2MB 25 cycle hit latency	N/A	8 way 2MB 16 cycle hit latency
Memory bus width	64 bit @ 400MHz / 1024 bit 250MHz	1024 bit 250MHz	64 bit @ 333MHz
NIC location	PCIBus	Memory Bus	Memory Bus

Overall Network Bandwidth – Mbps



Energy Efficiency Pareto Chart



Specweb99