



Technology Trends for Petascale Computing

Arndt Bode

Vice-President and CIO, Technische Universität München

Member of the Board of Directors, LRZ - Bavarian Academy of Sciences

DEISA Symposium

Towards Petascale Computing in Europe

22 Mai 2007



Technology Trends for Petascale Computing

- Architectures: Multicore – Accelerators – Special Purpose –
Reconfigurable – Memory and Cache –
Network – Secondary Storage
- Computing and Software: Programming Models – Heterogeneous vs.
Homogeneous – Compilers and Threading –
Operating Systems – Tools – Adaptivity
- Applications: Scalability



Trends in Microprocessor Technology

- Compute Performance of COTS Microprocessors depends of
 - Clock Frequency
 - Computer Organization

- Clock Frequency:
 - Exponential Increase impossible (Power, Cooling, 135 W/Chip)
 - Partly Solutions: Clock- and Power reduction, Sleep Transistors, ...
 - New Goal for Optimization: Energy-Efficiency: MIPS, FLOPS per W

- Computer Organization: ILP, Processor-internal Organization fully exploited
 - Pipelining, Superscalar, VLIW, Wordlength, ..., contra-productive to Energy-Efficiency (Speculation)
 - Future is Processor/Thread level-Parallelism
 - Multi-Core and Application-specific Processors , Fault Tolerance



Energy - Efficiency

$$P \sim A C V^2 f$$

P: Power

A: Activity Factor (Active Transistors on Chip)

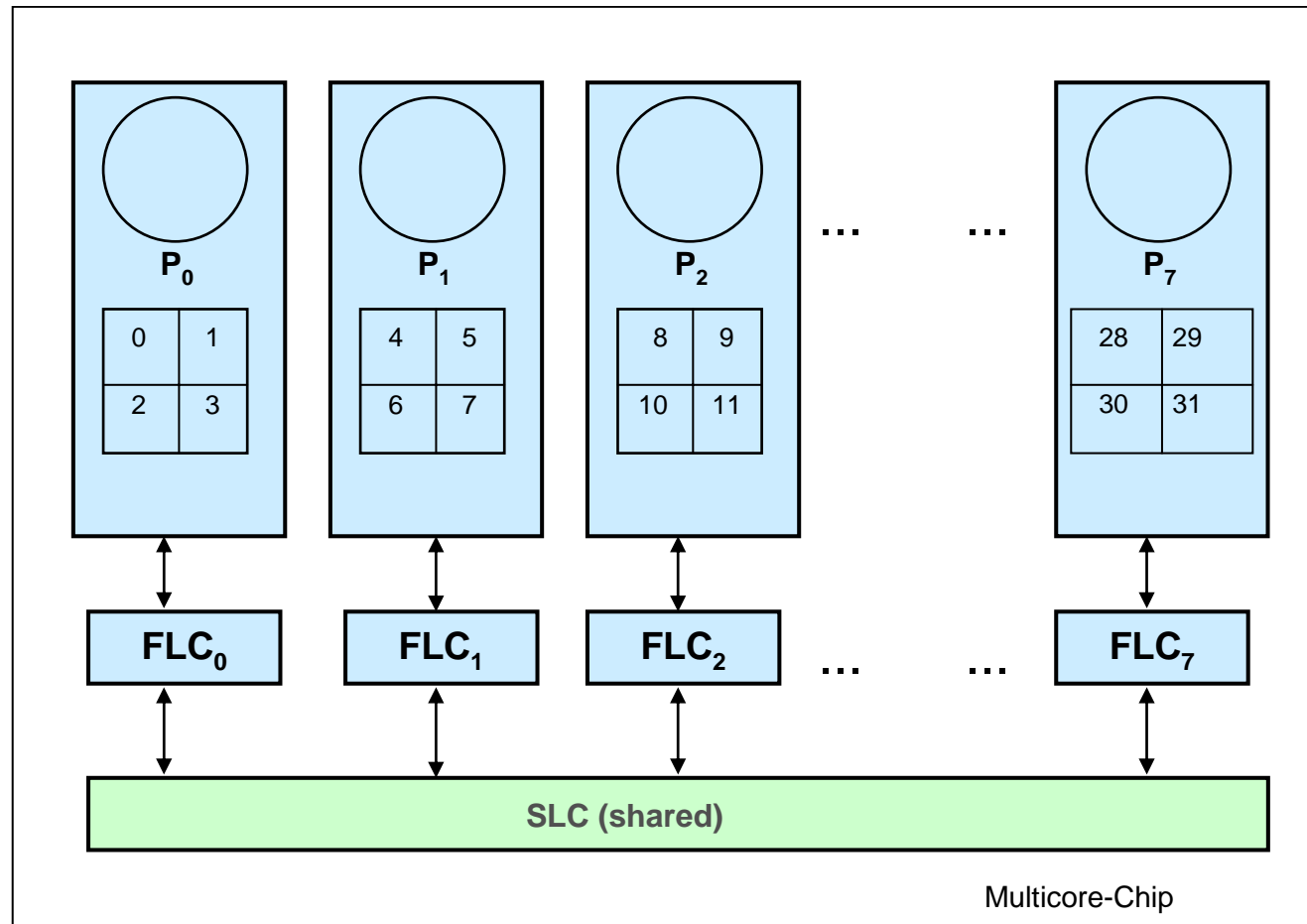
C: Total Capacity

V: Voltage

F: Clock Frequency

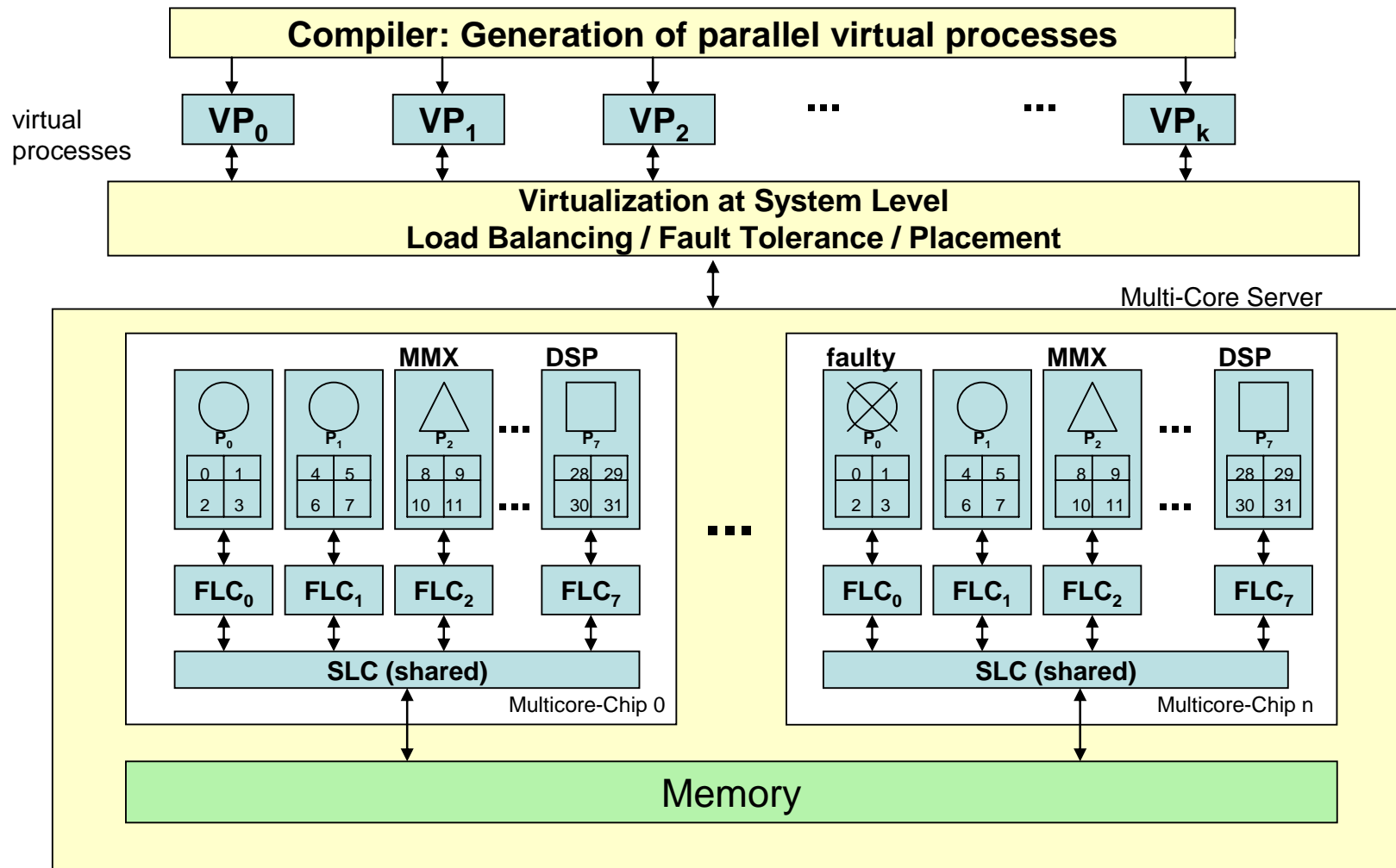


Multi-Core and Multithreading





Multi-Core Architectures: A Parallel World for Everybody





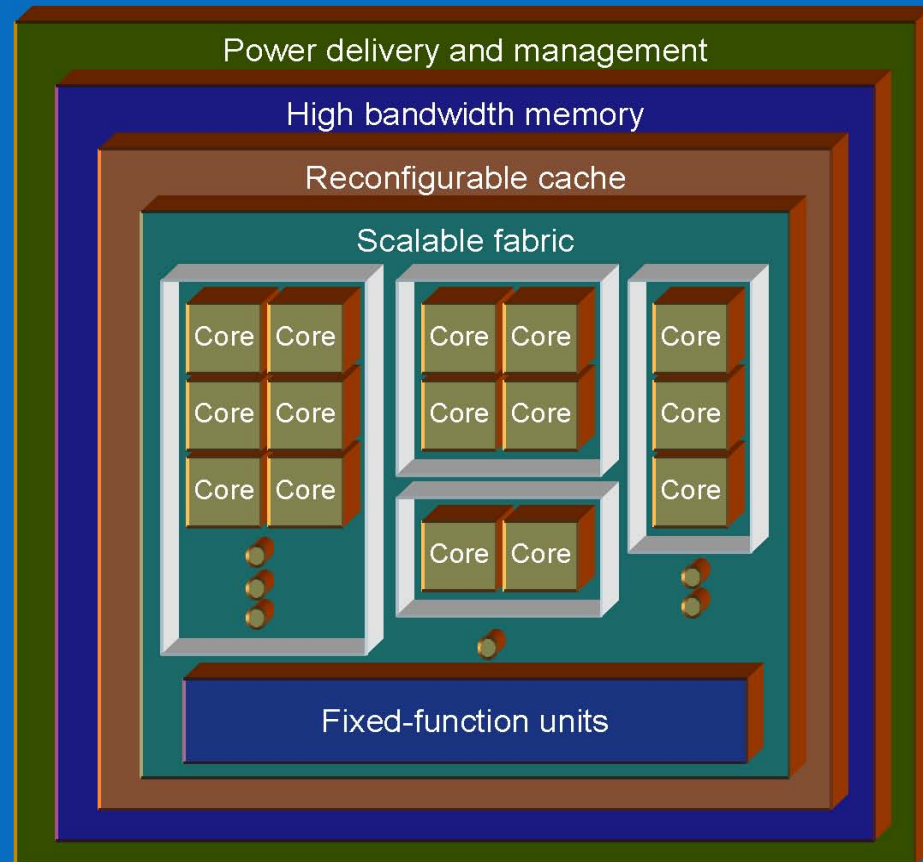
Platform Architecture

Parallel extension of IA

- Homogeneous array of cores
- Fixed-function units
- Coarse- and fine-grained data- and thread-level parallelism
- Global coherency hardware

Partitioned array

- Application domains
- Isolated communication traffic
- Fault tolerance





Processor-Design Options for Petaflop Systems

- Mega-Multicore Systems: General Purpose Applications (Itanium, Xeon, Power-7, ...)
- Multicore and Attached Accelerators (Cell, ClearSpeed, ...)
- Special Purpose (BlueGene, QCD, POLARIS, ...)
- Reconfigurable (FPGA: continuous research)

General Purpose vs. Special Purpose

- Programmability
- Scalability
- Applicability

- High Performance
- Low Power

History: Special Purpose „volatile“, Programming Interface not Compatible
Microprogrammable Devices : Vertical Migration and Bitslice
Early Microprocessors : Coprocessors (FP, IO, DSP, ...)
Early HPC : Vector Processor Attachments, Array-Processors, Associative Processors



Petaflop Processor Options: An Oracle

Two Types of Systems will persist:

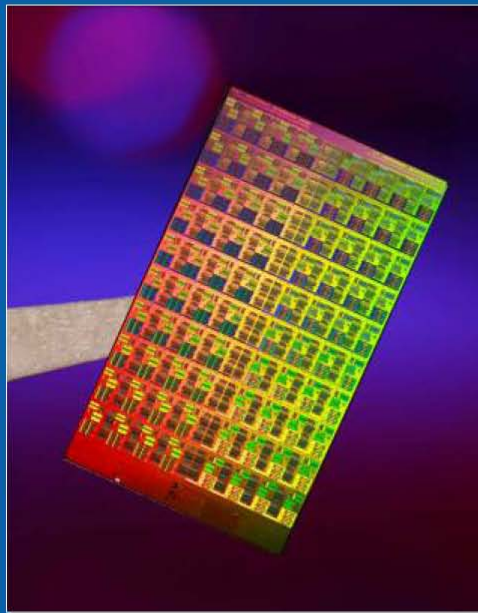
- General Purpose Mega-Multicore (Programmability, Compatibility, Cost)
- Special Purpose Systems for Large Specific Application Classes (Energy Efficiency)

The Systems will be Highly Parallel



Teraflops Research Chip

100 Million Transistors • 80 Tiles • 275mm²



First tera-scale programmable silicon:

- Teraflops performance
- Tile design approach
- On-die mesh network
- Novel clocking
- Power-aware capability
- Supports 3D-memory

Not designed for IA or product

Content under media embargo through
Sunday, February 11th Noon PST



8



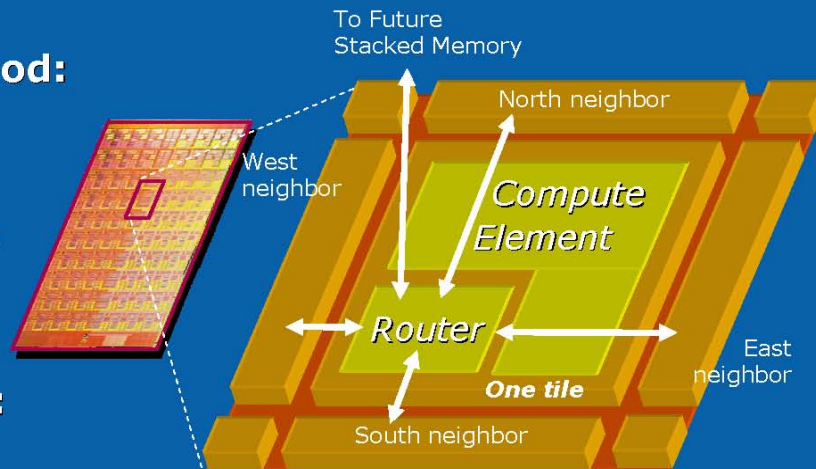
Tiled Design & Mesh Network

Repeated Tile Method:

- Compute + router
- Modular, scalable
- Small design teams
- Short design cycle

Mesh Interconnect:

- "Network-on-a-Chip"
 - Cores networked in a grid allows for super high bandwidth communications in and between cores
- 5-port, 80GB/s* routers
- Low latency (1.25ns*)
- Future: connect IA/or and special purpose cores



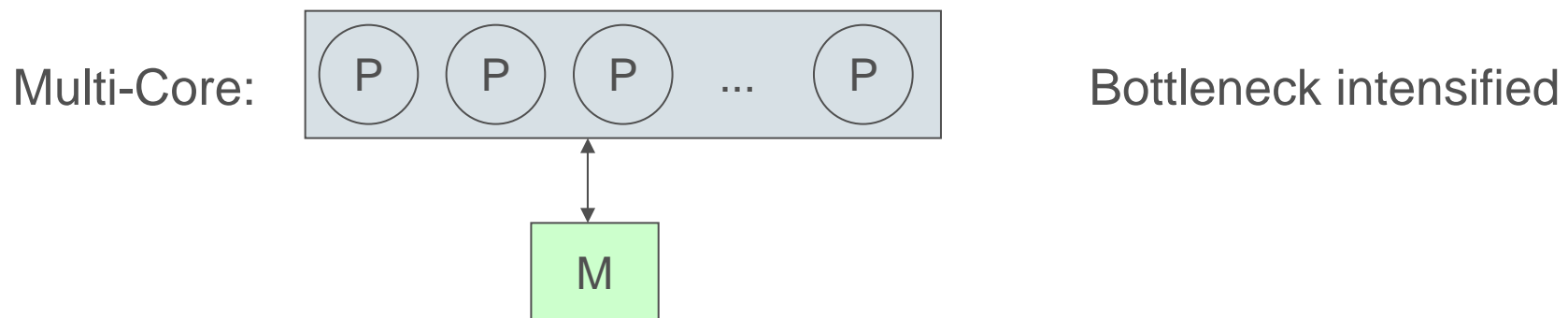
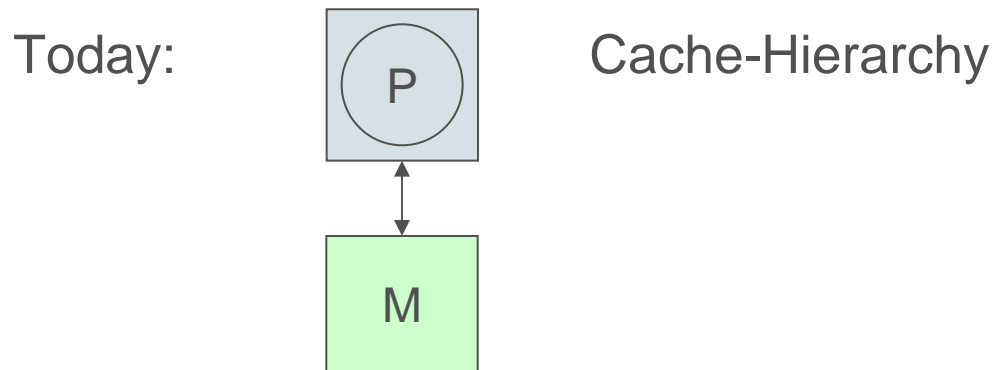
* When operating at a nominal speed of 4GHz
Content under media embargo through
Sunday, February 11th Noon PST





Consequences for Computer Architecture

Memory-Bottleneck



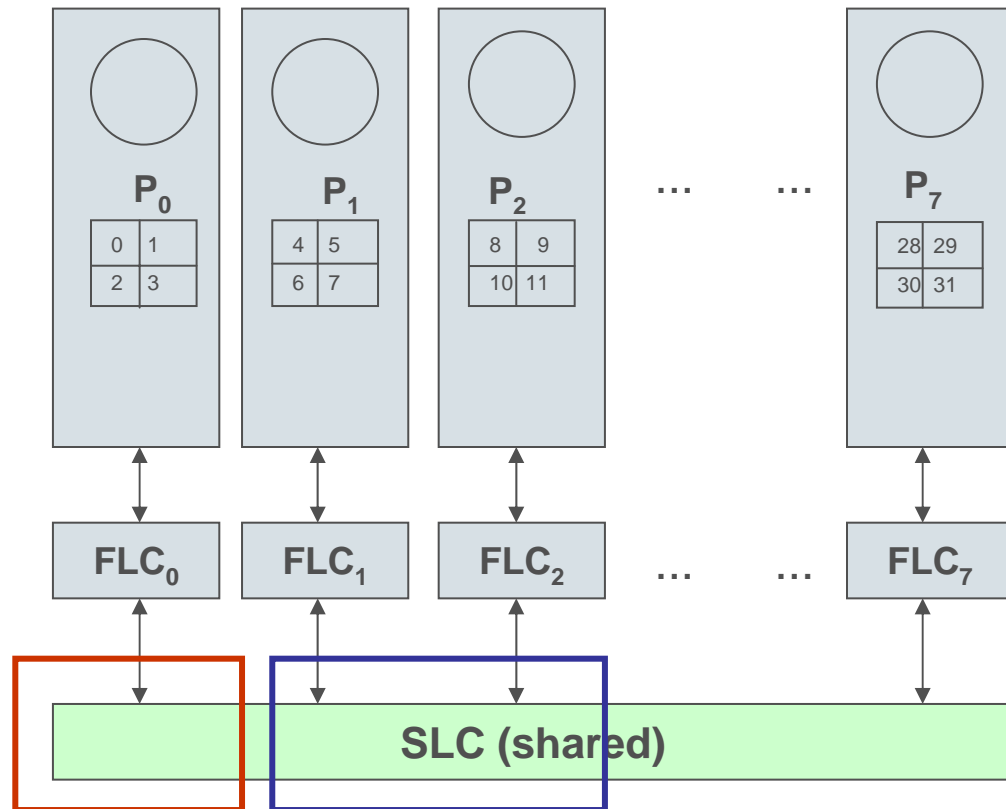


Cache Organizations

Distributed

Shared

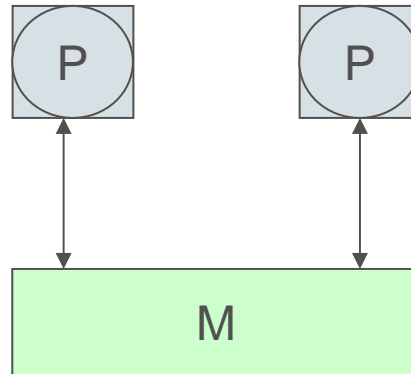
Groupwise shared





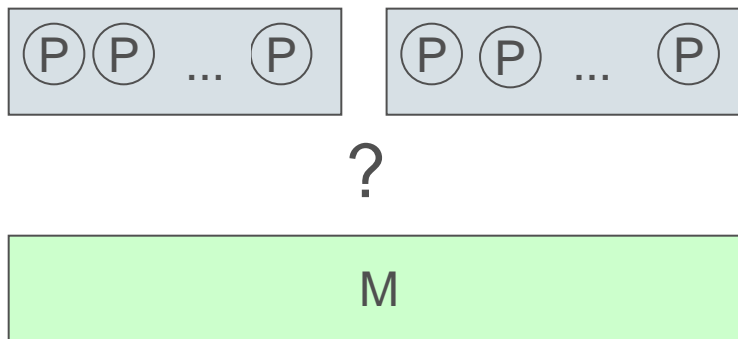
Communication Bottleneck

classical SMP:



1 Interface / Proc. IC

SMP based on Multi-Core



n Interfaces / Multi-Core IC

(e. g.: n cores
5n Interfaces
mesh-structure)

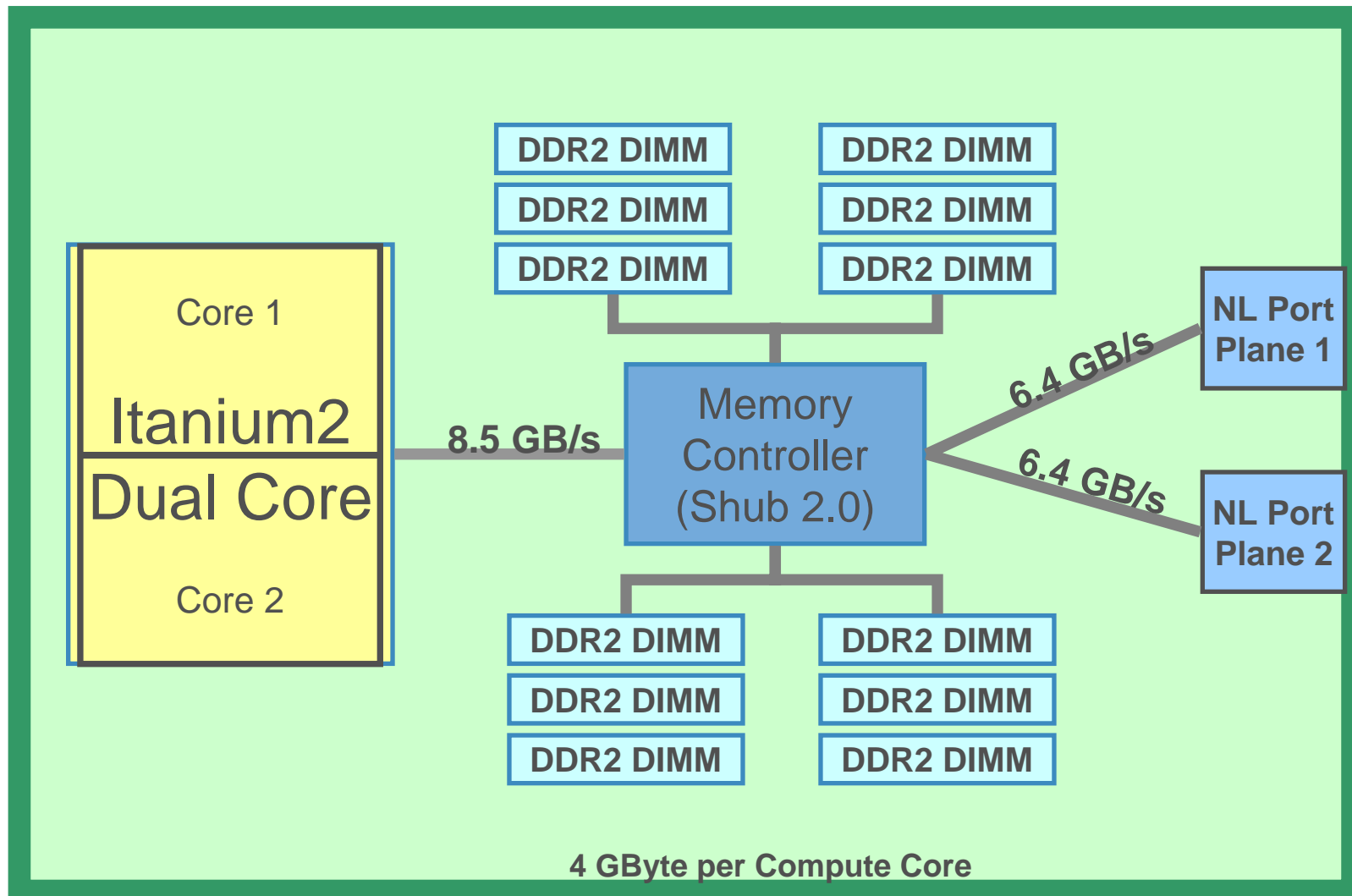


HLRB II: SGI ALTix 4700 / 9728 Montecito cores



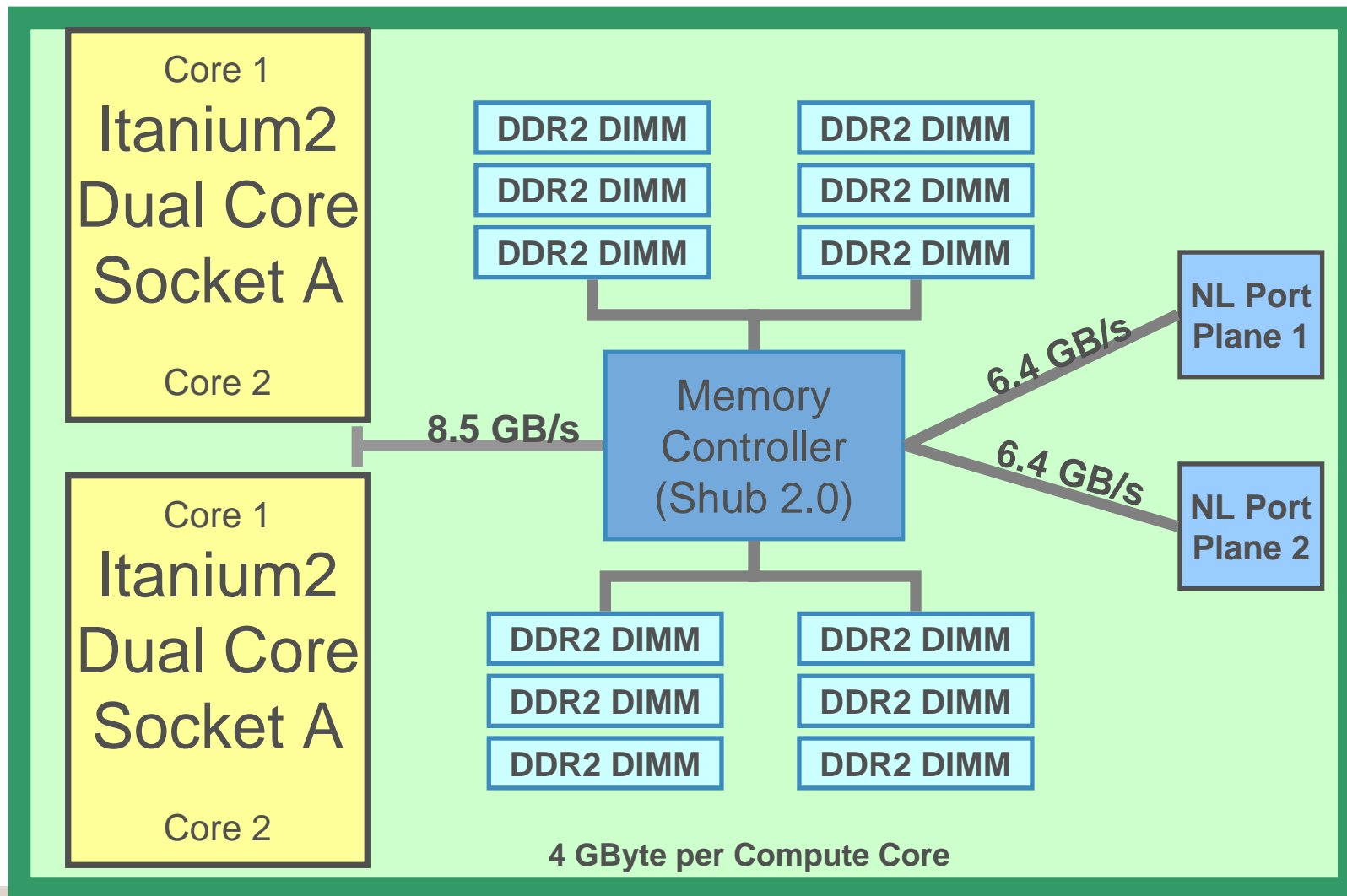


Blade





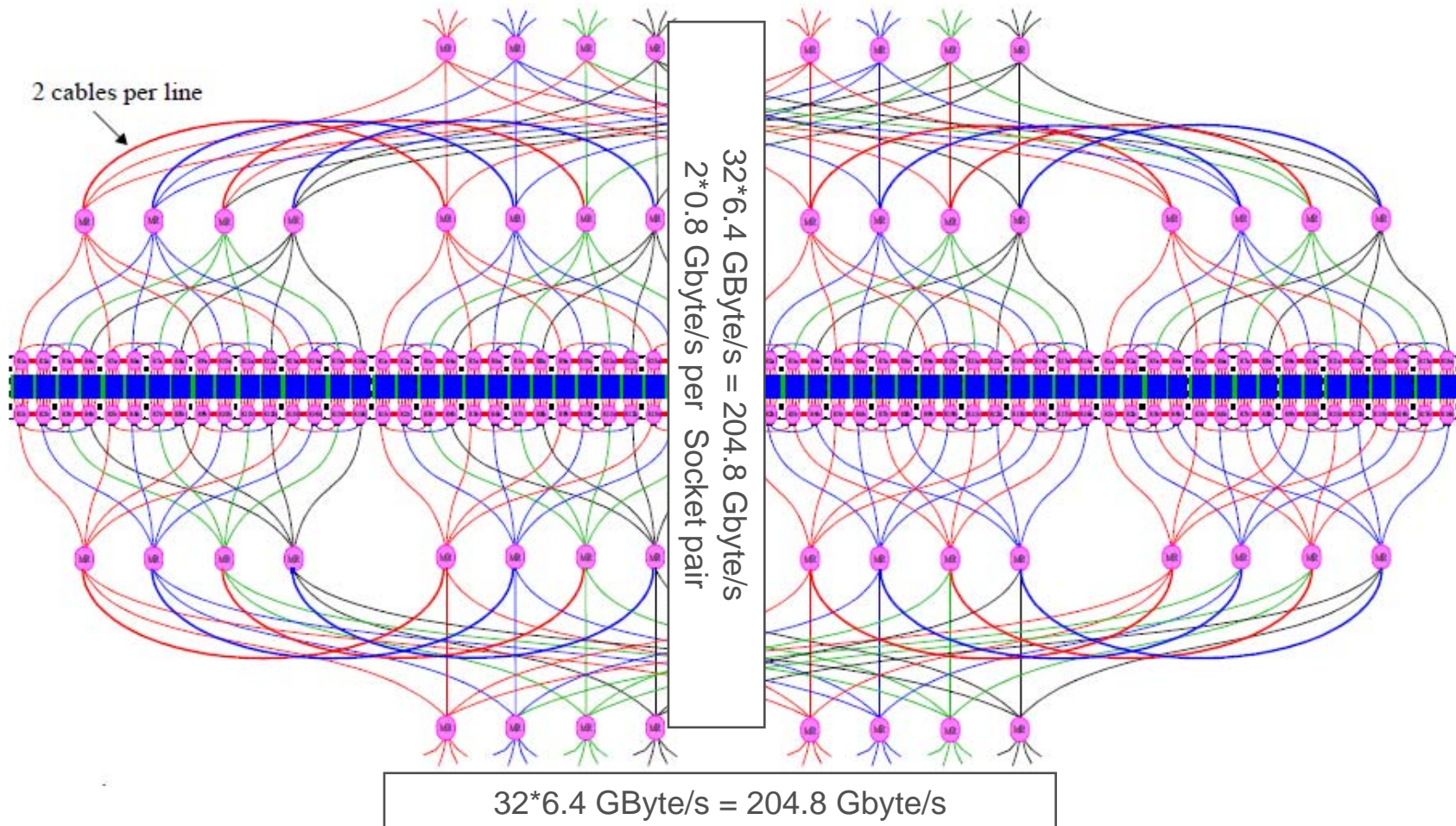
Dual Socket Blades - Density Compute Blades





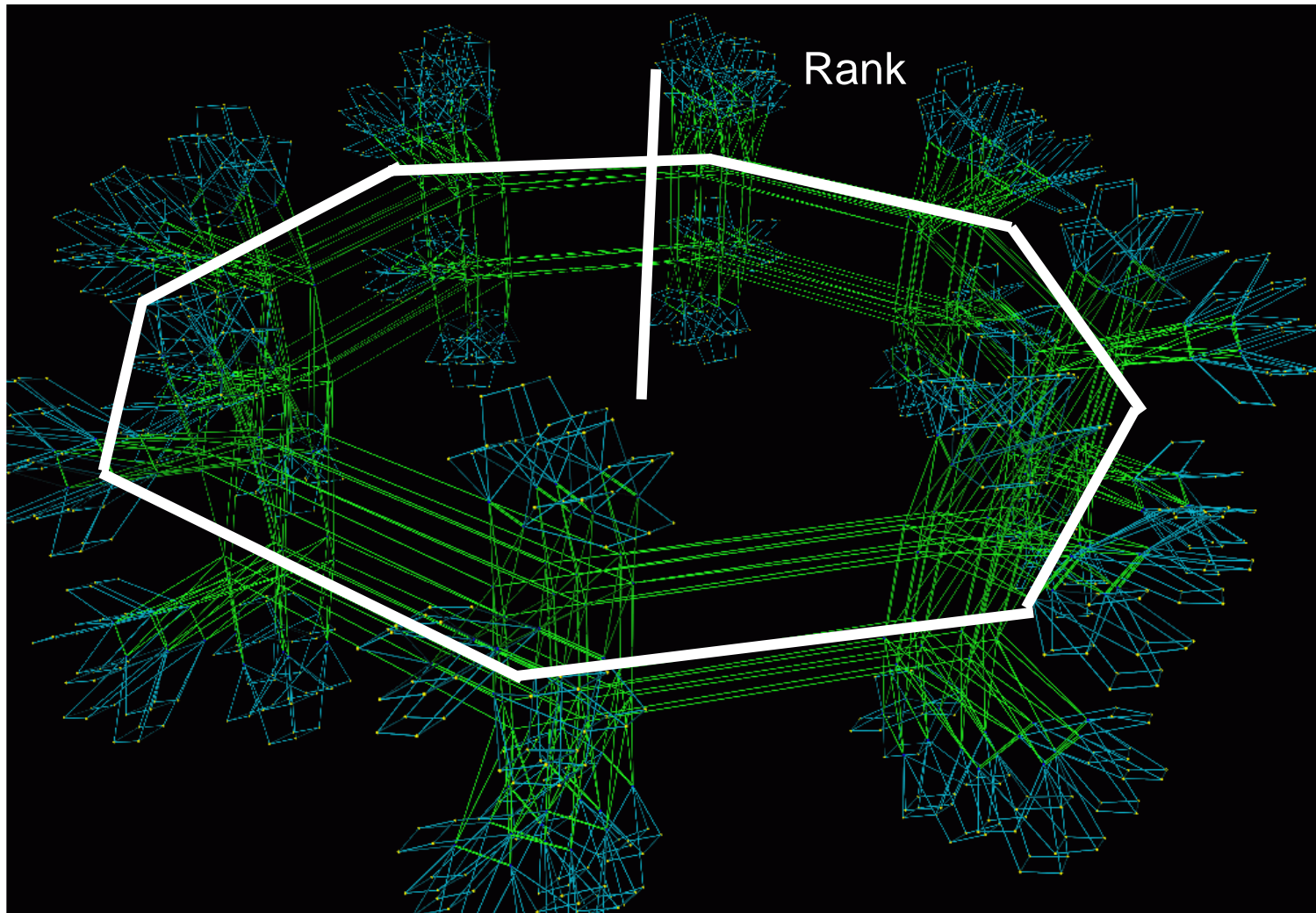
One 256 Socket Partition: ccNUMA Shared Memory

$32 * 6.4 \text{ GByte/s} = 204.8 \text{ Gbyte/s}$



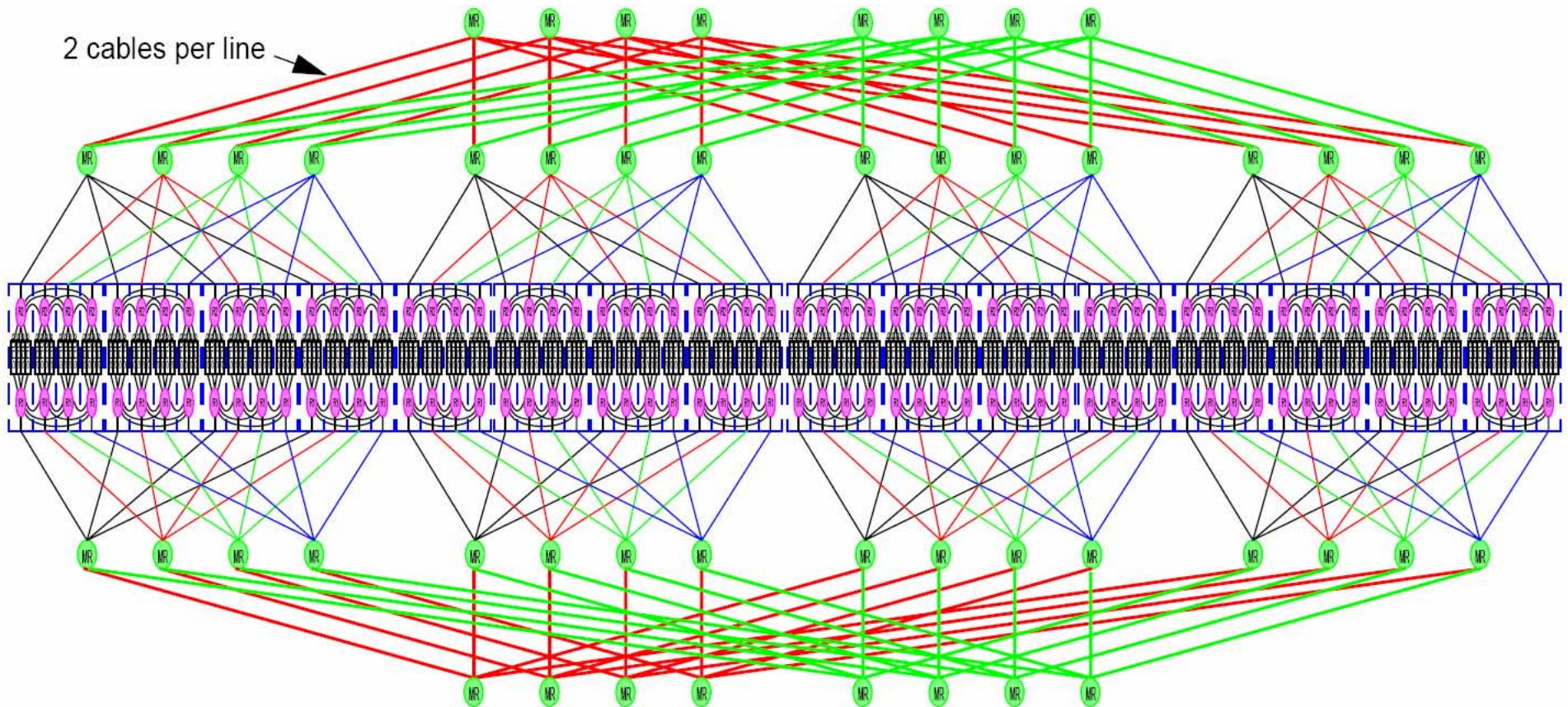


2-D Torus Mesh topology.





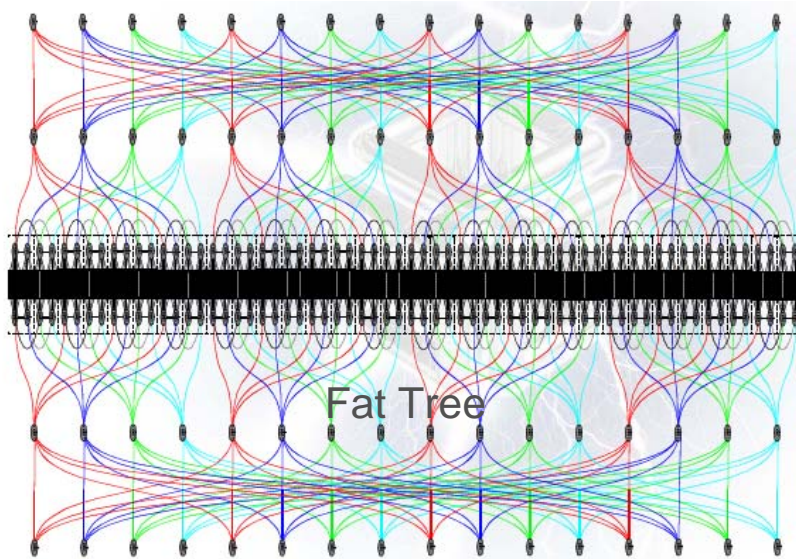
HLRB II Interconnect



Fat Tree Topology



Topology

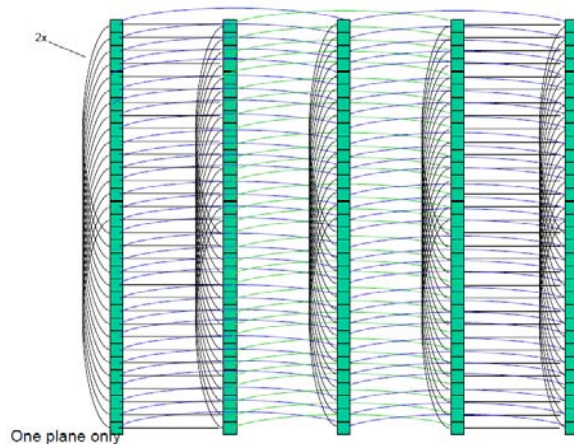


Batch Scheduler must be topology aware

PBSpro 8.0 early access

Placements of jobs is optimized according to topology

Topology is stored in placement sets



2D-Torus



Petascale Software

Key Issue is use of Excess Parallelism:

■ Programming Model, Language and Compiler

- Conventional
- Parallelizing (Including JIT)
- Parallelizing with Directives
- New Parallel Languages (Transactional Memory, ...)

■ Use of Threads

- Functional
- Speculative
- Assist/Helper (Prefetching, Monitoring, Debugging, Tools, Virtualization, Security, FT-Lockstep, ...)

■ Scalable Tool Models



Some Examples of Research at MMI



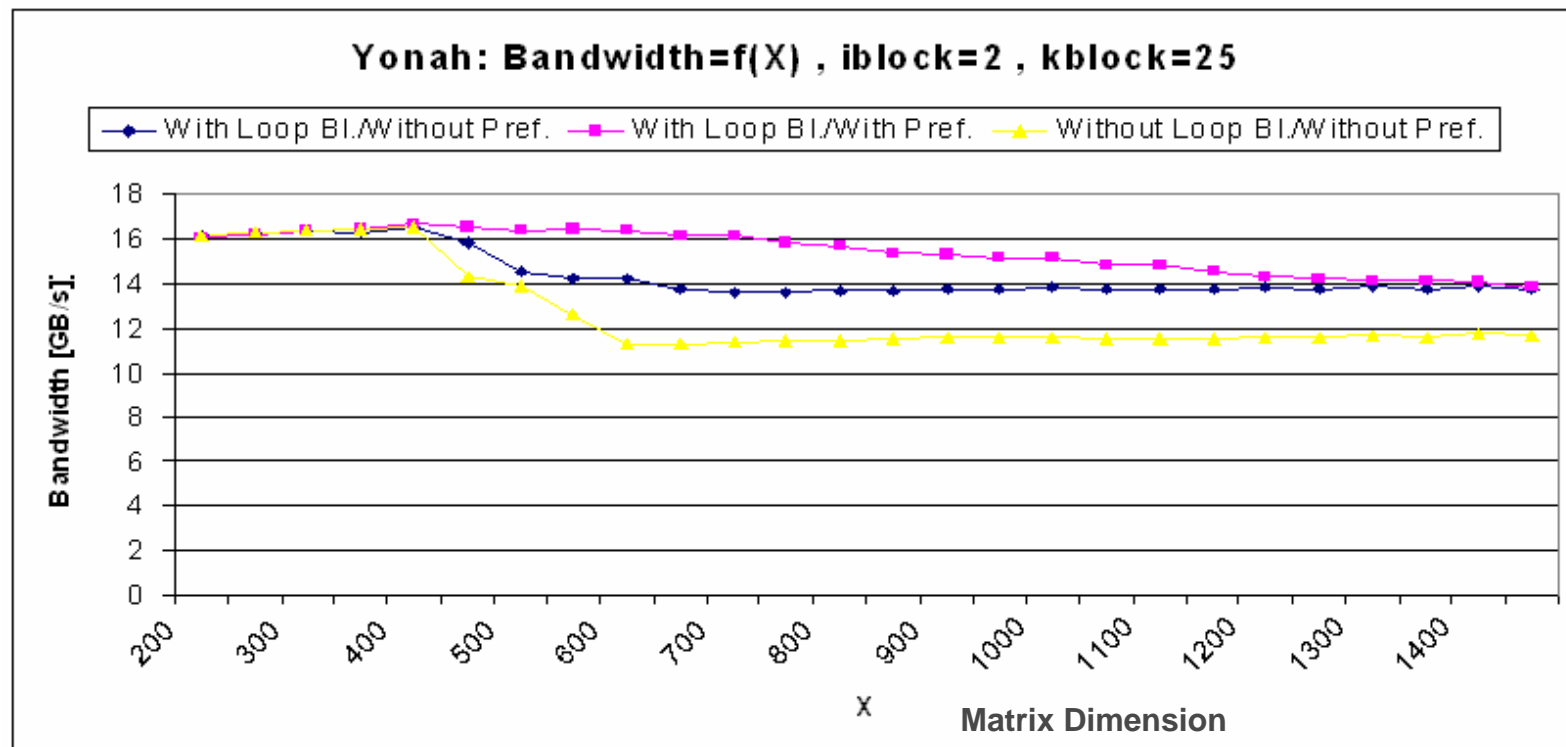
- Cache-Prefetching with Helper-Core (up to 39 %)
- Cache-Behavior with shared/separate Caches
- Helper Cores for Fault Tolerance



Prefetching

Bandwidth obtainable for Matrix-Multiplication

(Blocking with small amount of caching)

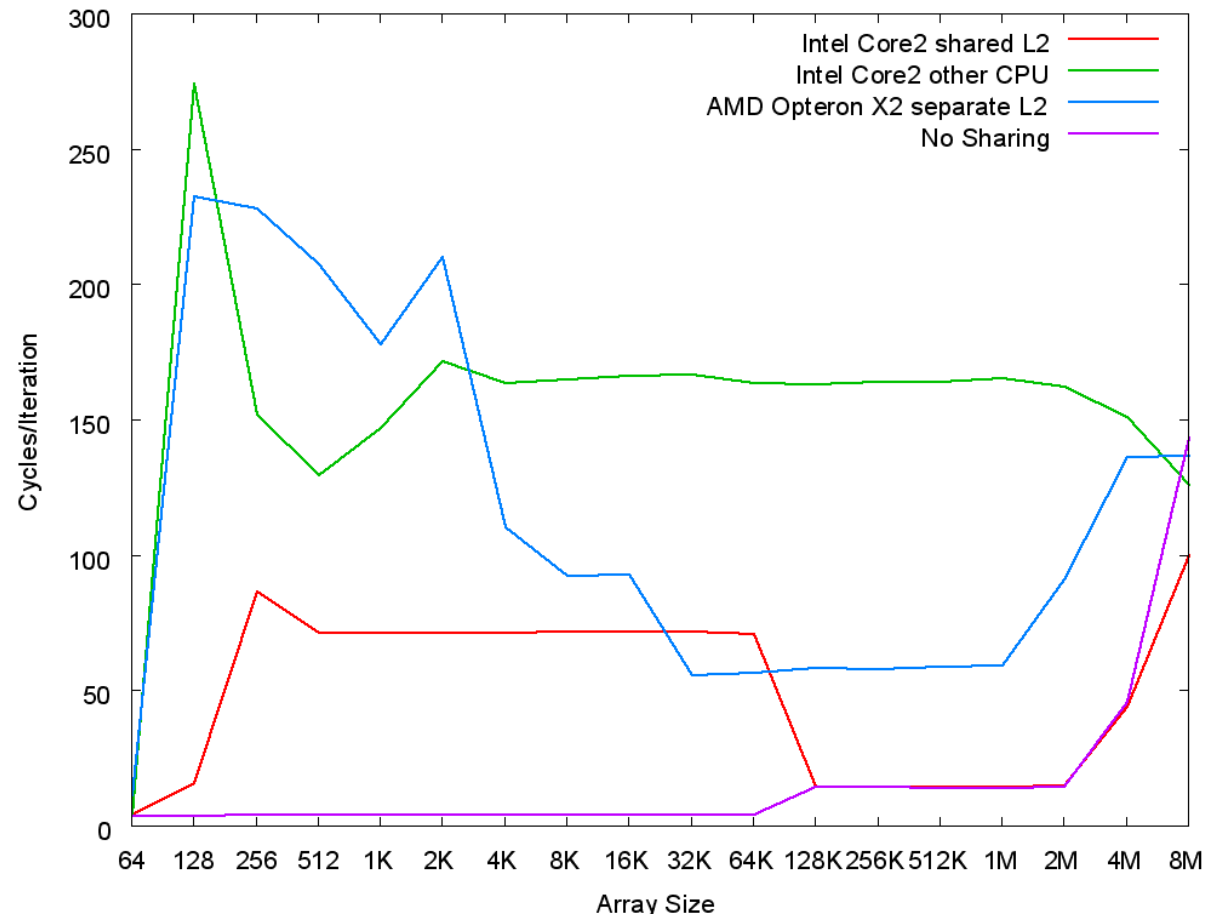




Advantages of Shared Caches

Synthetic Benchmark:

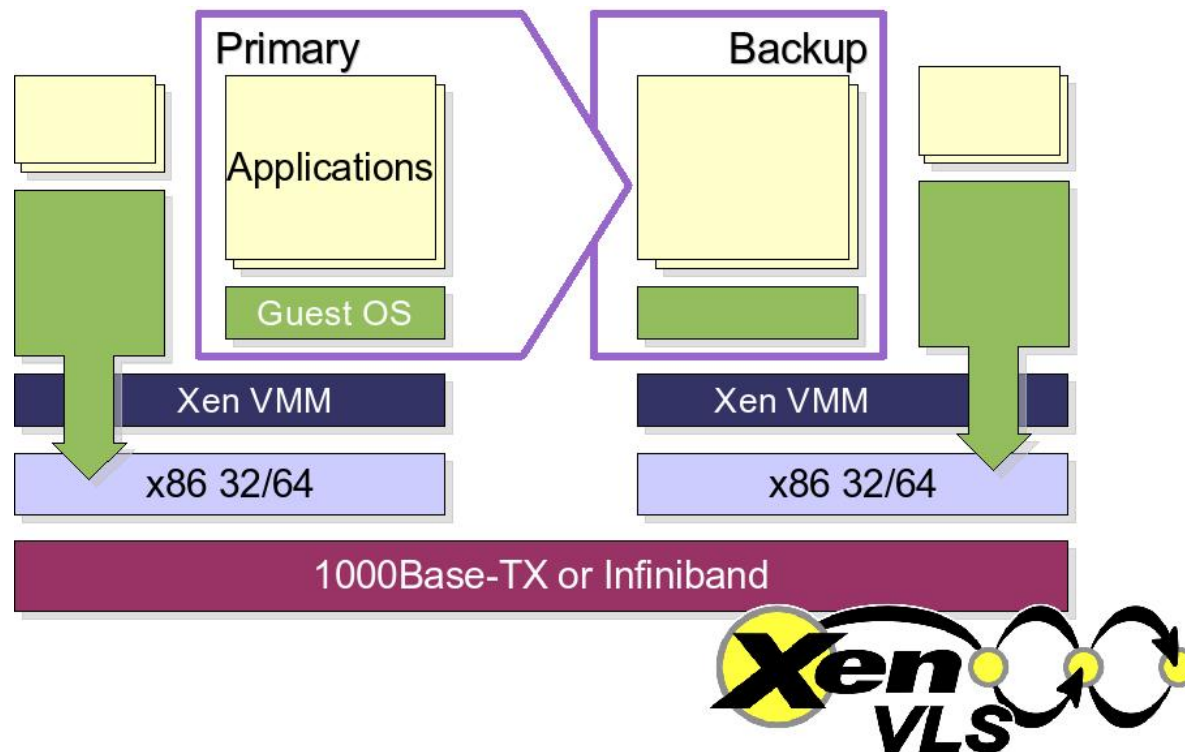
Latency for continuous Writes onto the same Memory area (2 cores)





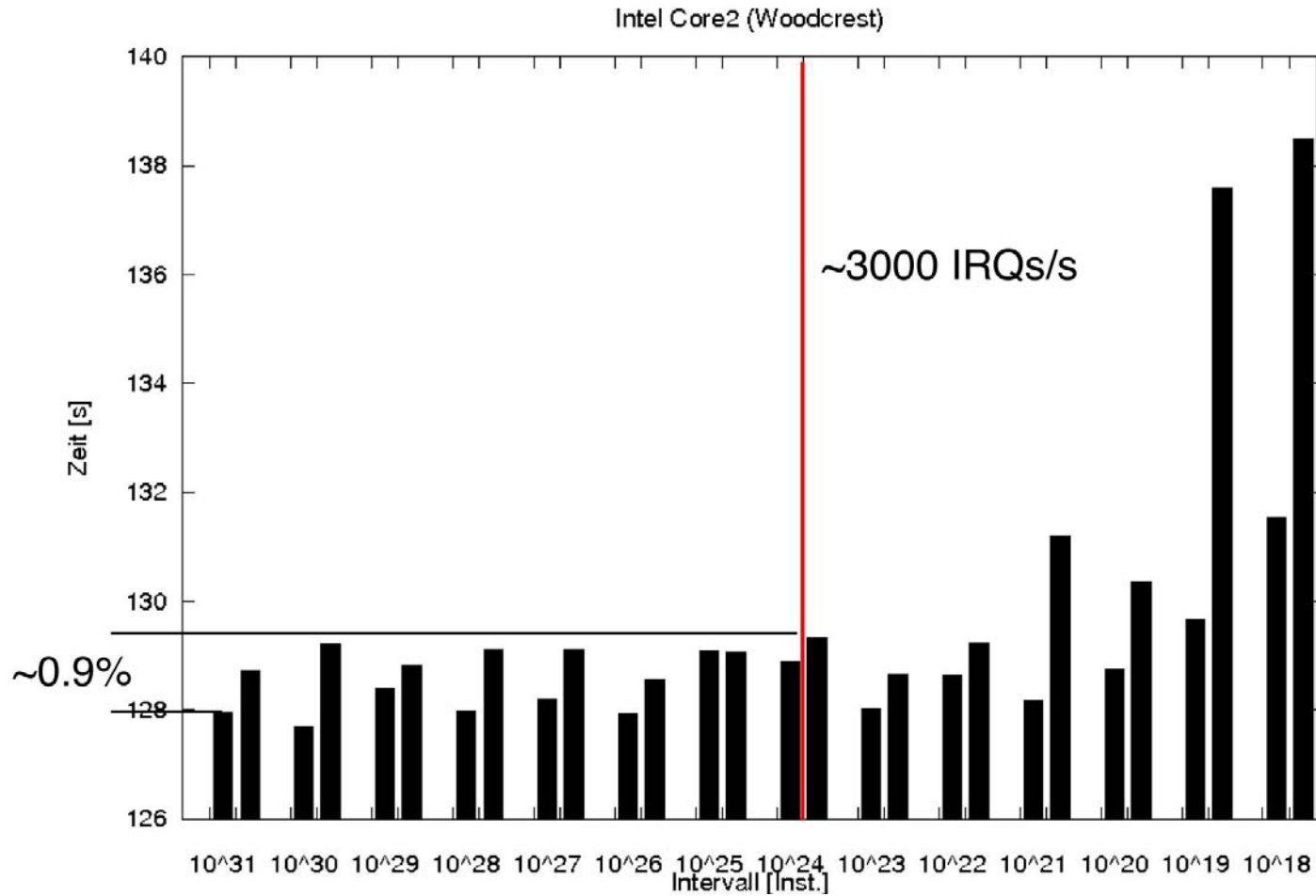
Lockstepping of Virtual Machines

- Synchronized Identical Execution of Applications on 2 Processors
- Virtual Environment allows for Monitoring and Control by Second Core
- Minimal Performance Loss





Lockstepping of Virtual Machines





Summary

- Many Choices in System Architecture and Programming Models
- More Research and Development Needed: DEISA and PACE
- Europe must have Access to General Purpose and Special Purpose Architectures



Munich's Contribution to Petaflop Computing

Infrastructure

LRZ: German and Bavarian Supercomputers
(SGI ALTIX 4700, ...)

RZG: Max Planck Supercomputer Center
(IBM pSeries Regatta, ...)

MCSC: = LRZ + RZG; LRZ member **GCS**

Research

KONWIHR: Bavarian Competence Network for HPC
includes TUM, LMU (Excellence-Universities)

DEISA, PACE, ...

D-Grid, LHC, ...

Teaching

BGCE: Bavarian Graduate School of Computational Engineering:
Master of Science with Honors (CSE, Comp. Mechanics, CE)

IGSSE: Intern. Graduate School for Science and Engineering

Industrial Applications

Joint Projects; BMW, Audi, EADS, Siemens, ...



Welcome to the Bavarian Graduate School of Computational Engineering

Computational Engineering ...

refers to all activities in engineering that use computers as their main tool. Typical tasks in Computational Engineering are the solution of differential equations that model certain physical phenomena, the optimization of processes in engineering, or the stochastic simulation of a complex system.

The Bavarian Graduate School ...

of Computational Engineering is an association of three Master programs:

- **Computational Engineering (CE)** at the Friedrich-Alexander-Universität Erlangen-Nürnberg,
- **Computational Mechanics (COME)**, and
- **Computational Science and Engineering (CSE)** at the Technische Universität München.

Our goal is to push forward the rapidly growing field of Computational Engineering by offering high-quality master programs for students who are interested in the field.



The Elitenetzwerk Bayern ...



is an initiative of the state of Bavaria to support the education and advancement of highly talented students. With the help of the Elitenetzwerk Bayern, we are able to offer an "elite" degree program for the best students in our master programs. Outstanding performance in one of the three Master's programs will be honoured by the newly formed academic degree of a "**Master of Science with Honours**".

For further information about these elite program, please read on in our section "**Information for students**".

IGSSE

BGCE is a partner of **IGSSE**, TUM's International Graduate School of Science and Engineering