

***Performance Evaluation
of SPEC OMP Benchmarks
on the InfiniBand-Based Cluster System***

**Shinyoung Kim, Seo Hee, Inho Park,
Prof. Seon Wook Kim
Department of Electronics and Computer Engineering
Korea University, Seoul, Korea
<http://compiler.korea.ac.kr>**

Outline



- Introduction

- InfiniBand-Based Distributed Virtual Shared-Memory Systems

- Performance Evaluation
 - Experiment Methods
 - Performance

- Conclusion

Introduction



- We evaluate performance of SPEC OMP benchmarks on the InfiniBand-based Distributed Virtual Shared-Memory(DVSM) system in detail.
- Our research helps the system and application developers to estimate the application of the OpenMP applications on a commodity cluster of workstation to use the state-of-the-art interconnection networks for the next decades

What is InfiniBand?

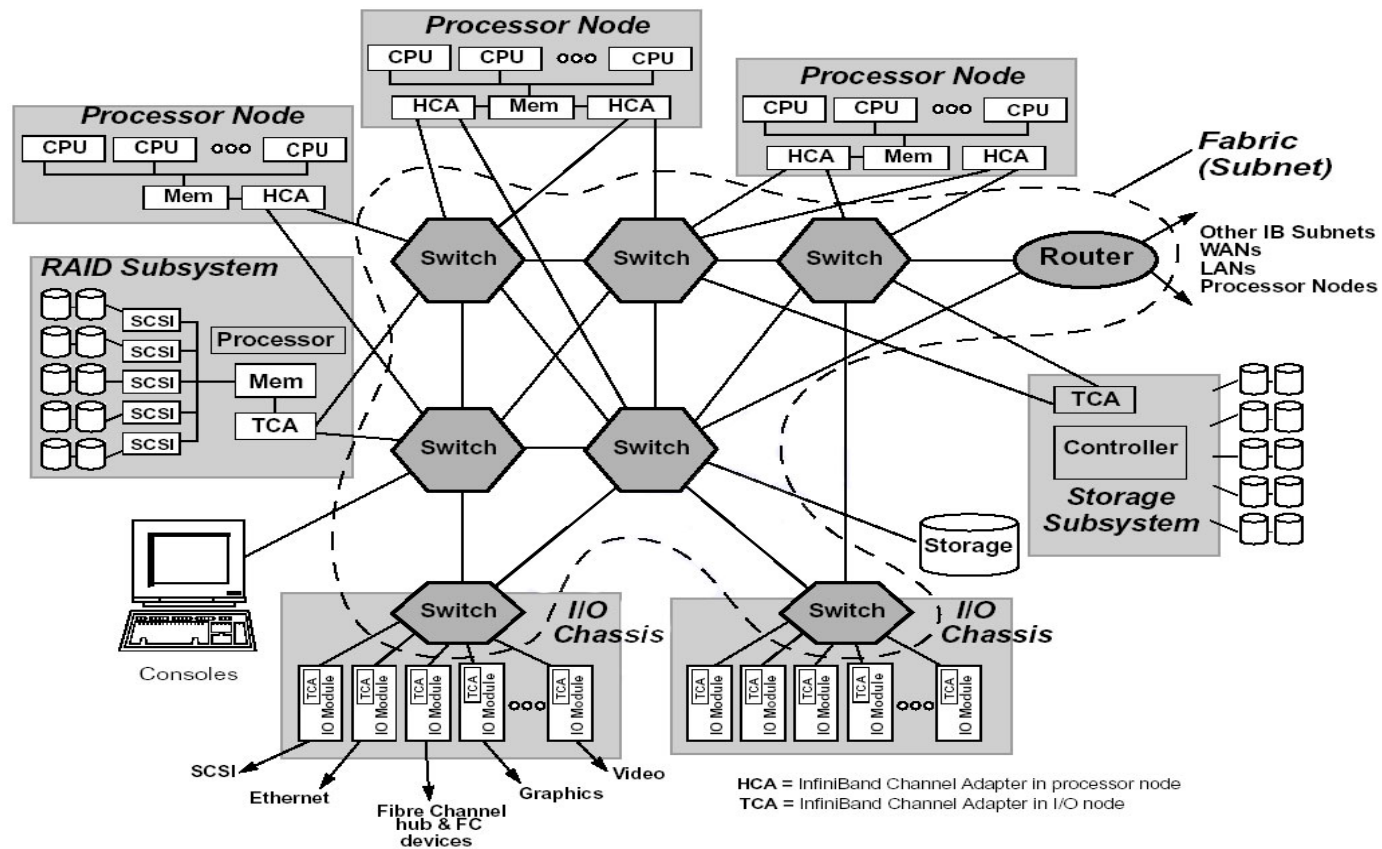


- The switched fabric in the IBA guarantees high stability and wider bandwidth.
- IBA does not need to interrupt other processes in order to access data on remote nodes because IBA provides hardware legacy software protocol tasks to support Remote Direct Memory Access (RDMA)
- IBA consists of processing nodes, I/O nodes, and System Area Network (SAN) to connect nodes.

InfiniBand Architecture



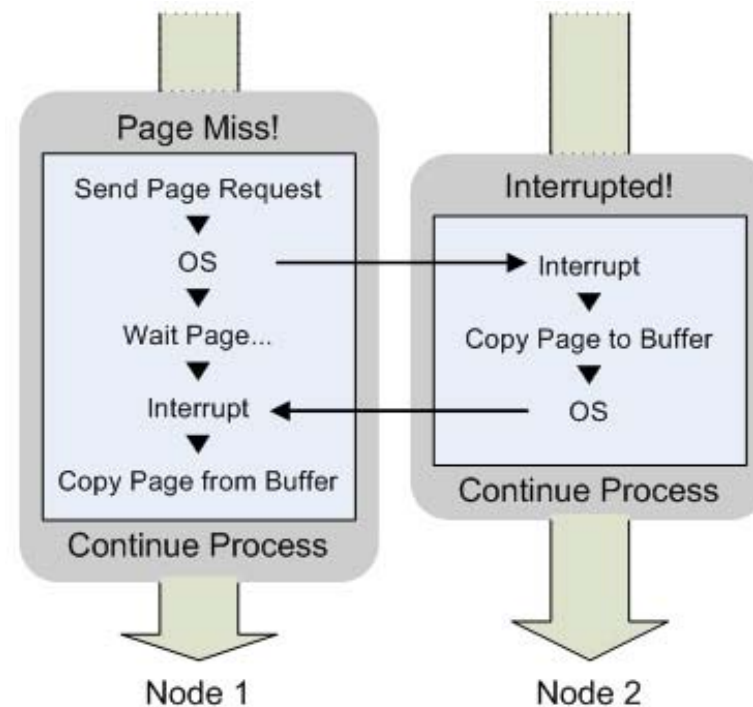
■ System Area Network (SAN)



Without RDMA



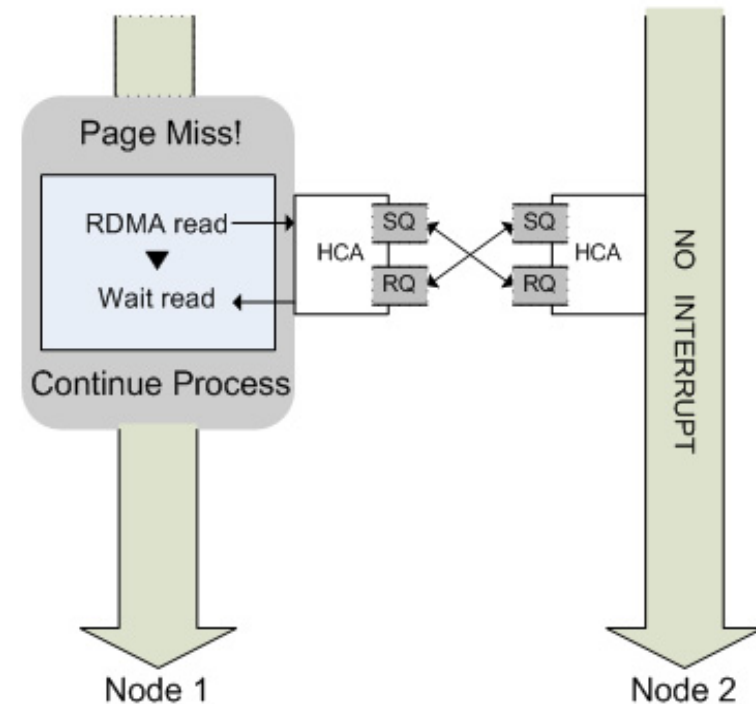
- Process should communicate directly with other processes in order to exchange data through process interruption



With RDMA



- Don't need to interrupt other processes in order to access data on remote nodes
- Reduce large execution overhead due to signal handling for process interruption and communication latency



InfiniBand-Based Distributed Virtual Shared-Memory Systems



- Our implementation is based on the lazy release consistency (LRC) model
- Because the RDMA feature removes process interruption for data communication, we need a new mechanism to track access history on the InfiniBand-Based DVSM system
- In order to record read and write operation per page, we use an interval table whose entry has two bits to mark read and write operations.

InfiniBand-Based Distributed Virtual Shared-Memory Systems



- Each synchronization includes two internal barriers.
 - At the first barrier each process broadcasts the interval table to all the participant processes
 - At the second barrier the diff operations are applied to other processes if necessary by considering all the other processes' interval tables.
- In order to reduce the occurrence of segmentation violations, we assign an ownership to each page.
 - Page owner is not memory-protected and it is write memory-protected only after other processes perform read and write operations.
 - The page owner information is maintained inside the interval table by extending one bit to mark a page ownership.

Experiment Methods



- Linux Cluster
 - Intel 2.0GHz Xeon processor, 512MB memory
 - 133MHz 4X PCI-X 128MB memory HCA board

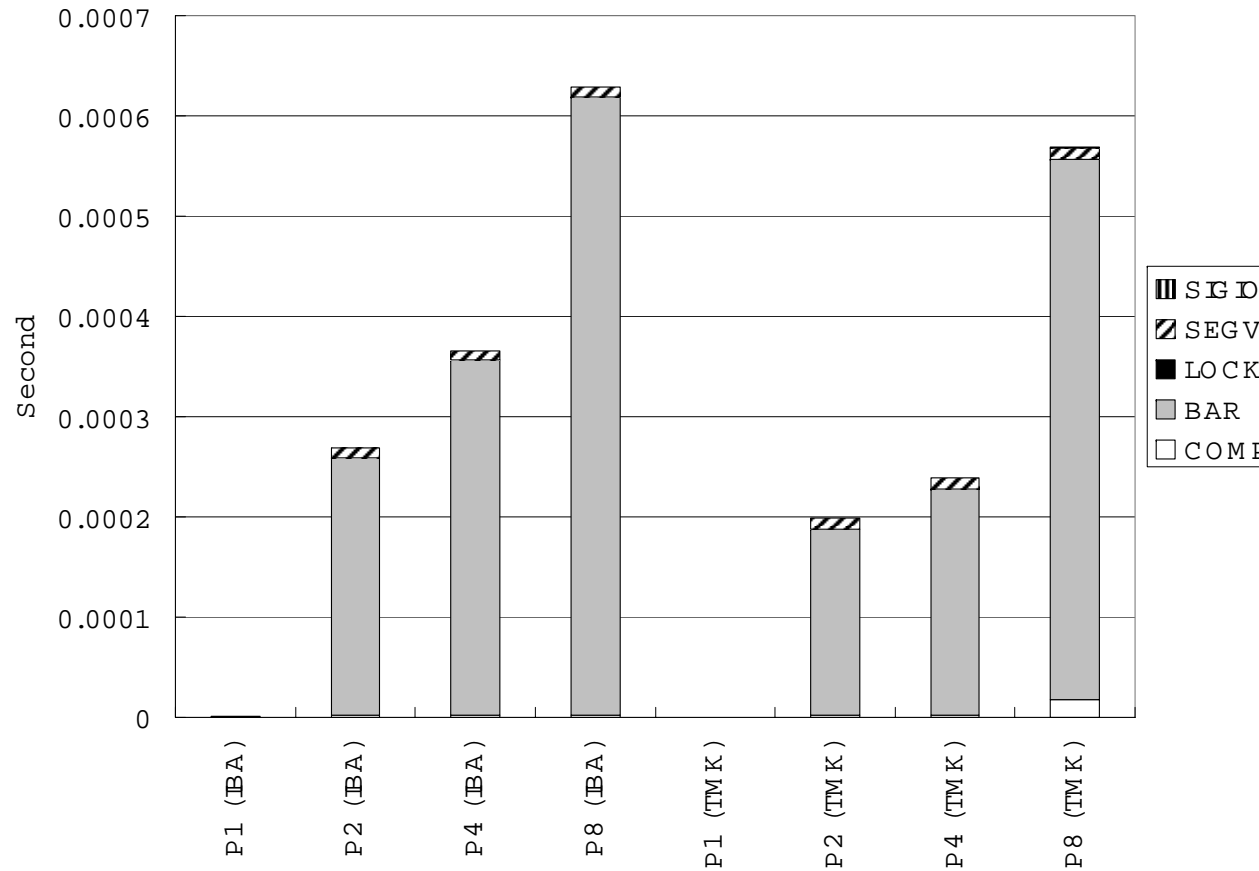
- SMP machine
 - Four 2.8GHz Xeon processors and 512MB memory

- Compiler
 - Polaris parallelizing compiler infrastructure

- Four benchmarks (swim, mgrid, wupise and applu)

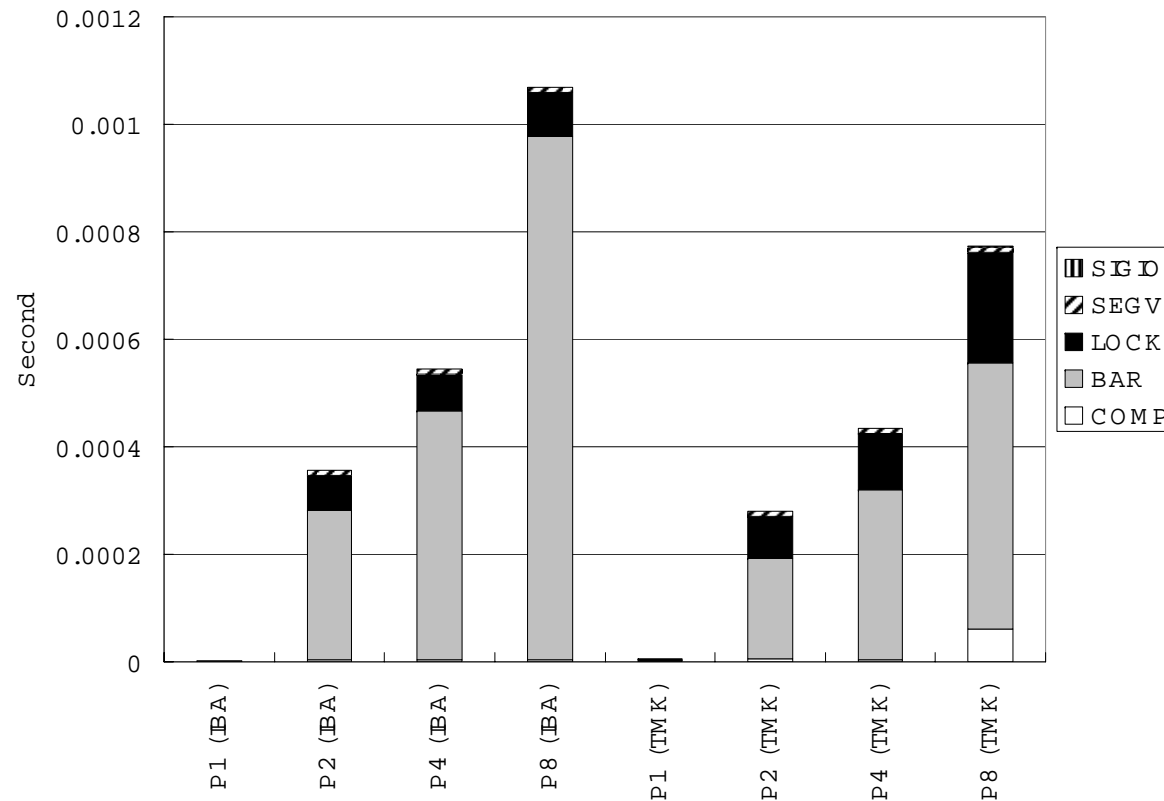
- Measurements for each benchmark
 - SMP – unoptimized
 - TreadMarks - unoptimized
 - InfiniBand-based DVSM – unoptimized, optimized

Overhead of OpenMP Primitives



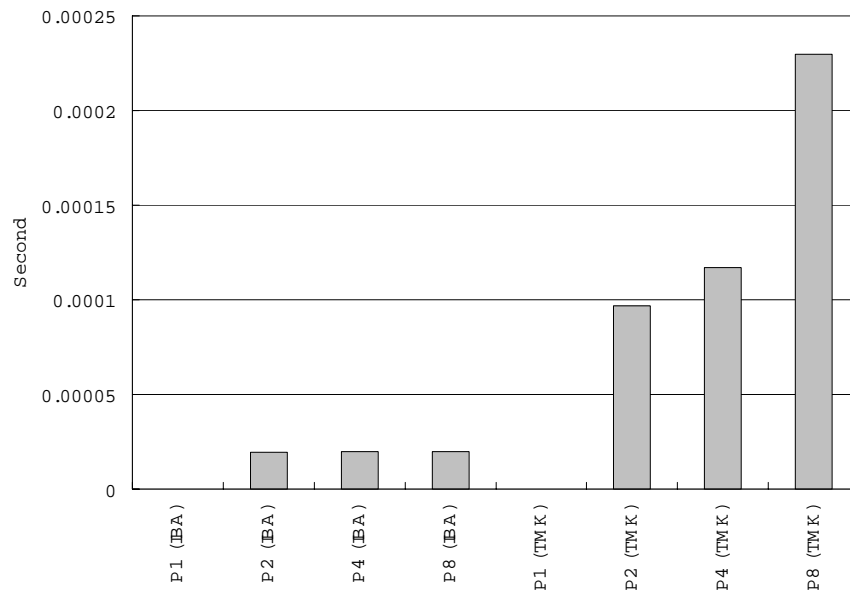
- Process fork/join overhead of the OpenMP parallel directive on our system (IBA) and TreadMarks (TMK)

Overhead of OpenMP Primitives

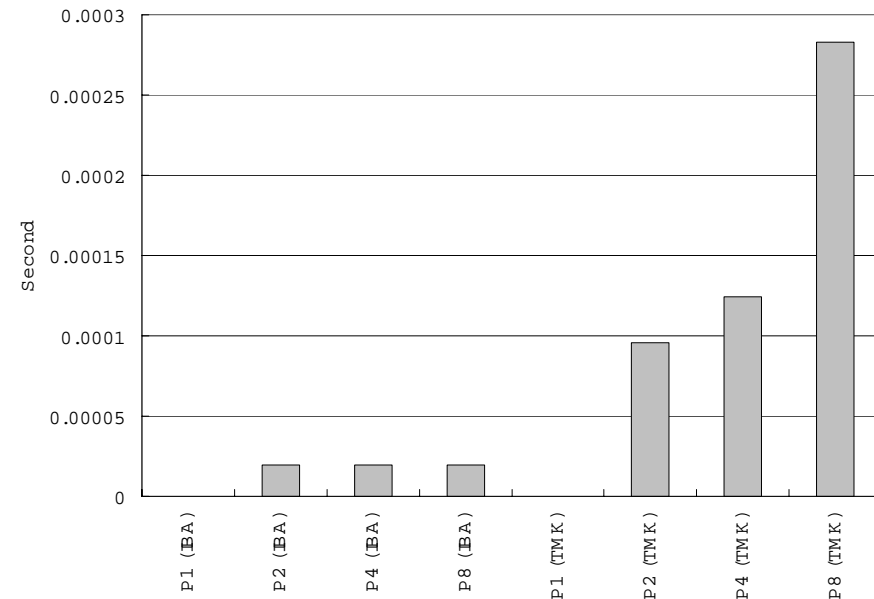


- Overhead of OpenMP atomic primitive on our system (IBA) and TreadMarks (TMK)

Overhead of page movement

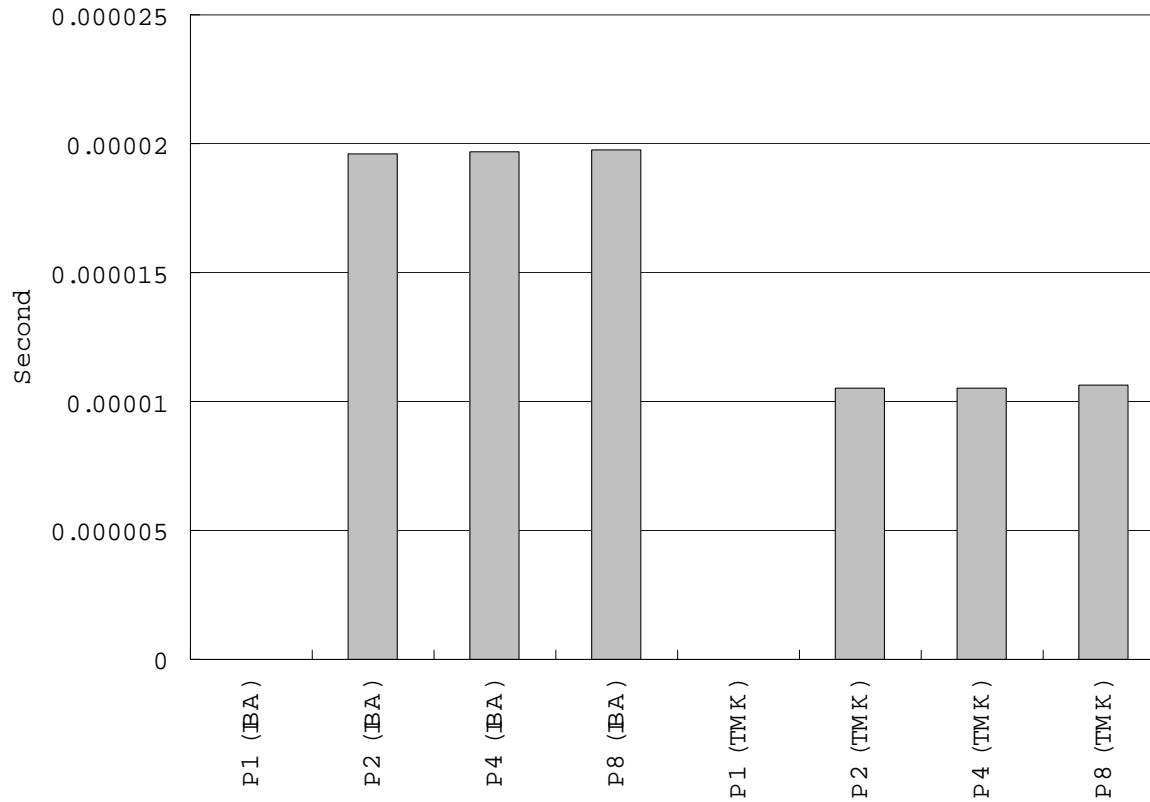


■ parallel to serial sections



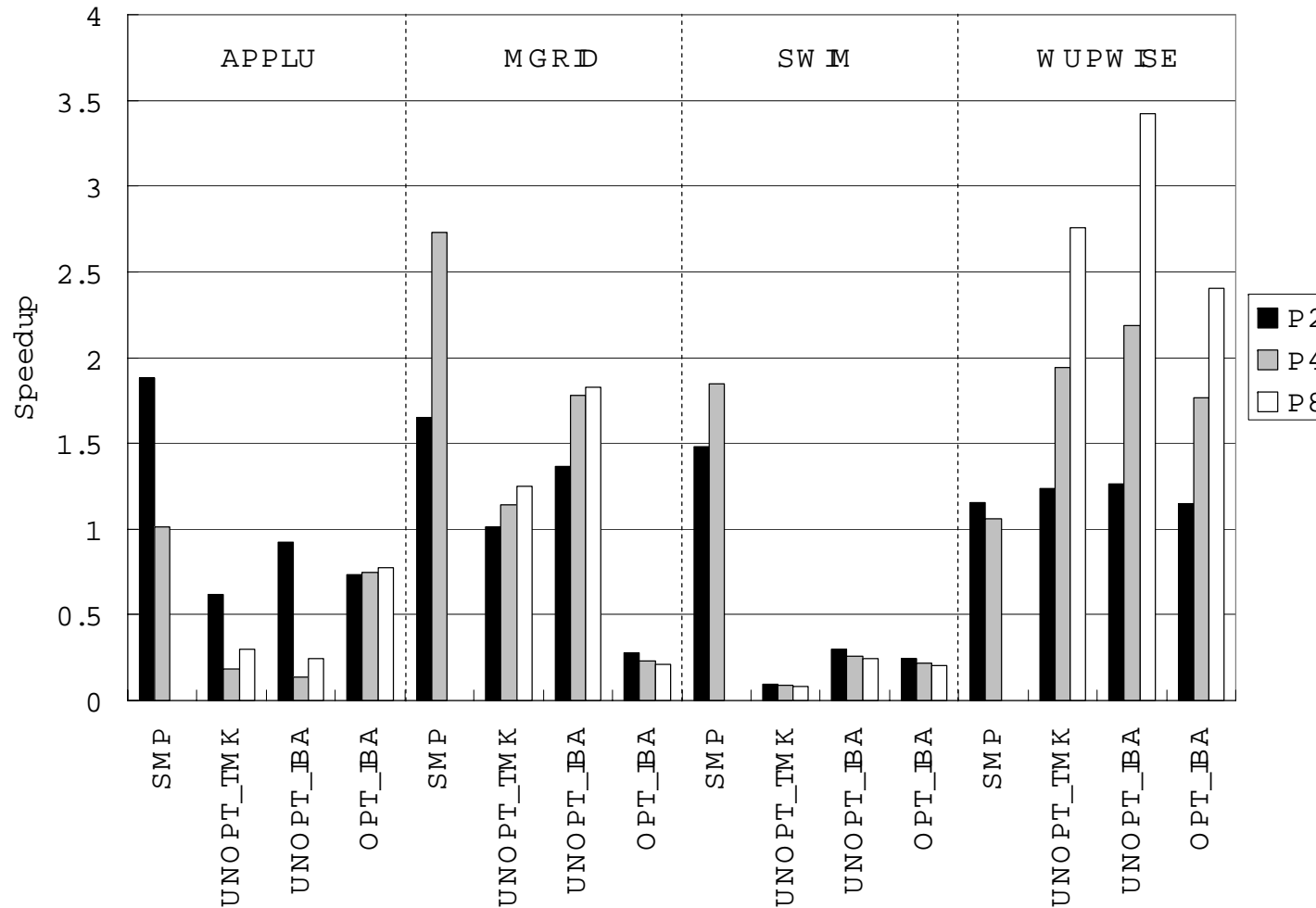
parallel to parallel sections

Overhead of page movement



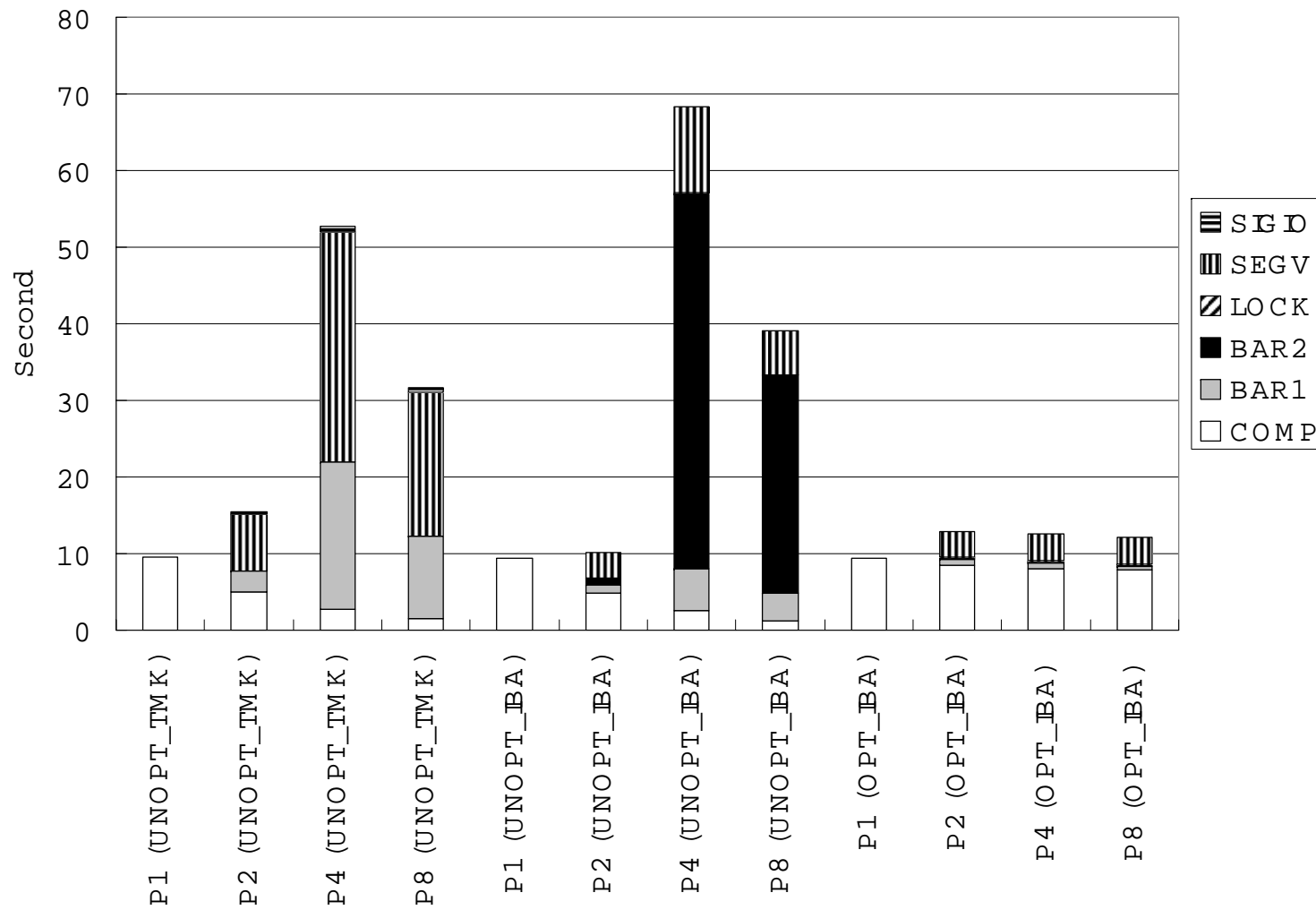
- Overhead of data movement from serial to parallel sections on our system (IBA) and TreadMarks (TMK)

Overall Performance of OMP Benchmarks



■ Speedup of the OpenMP applications

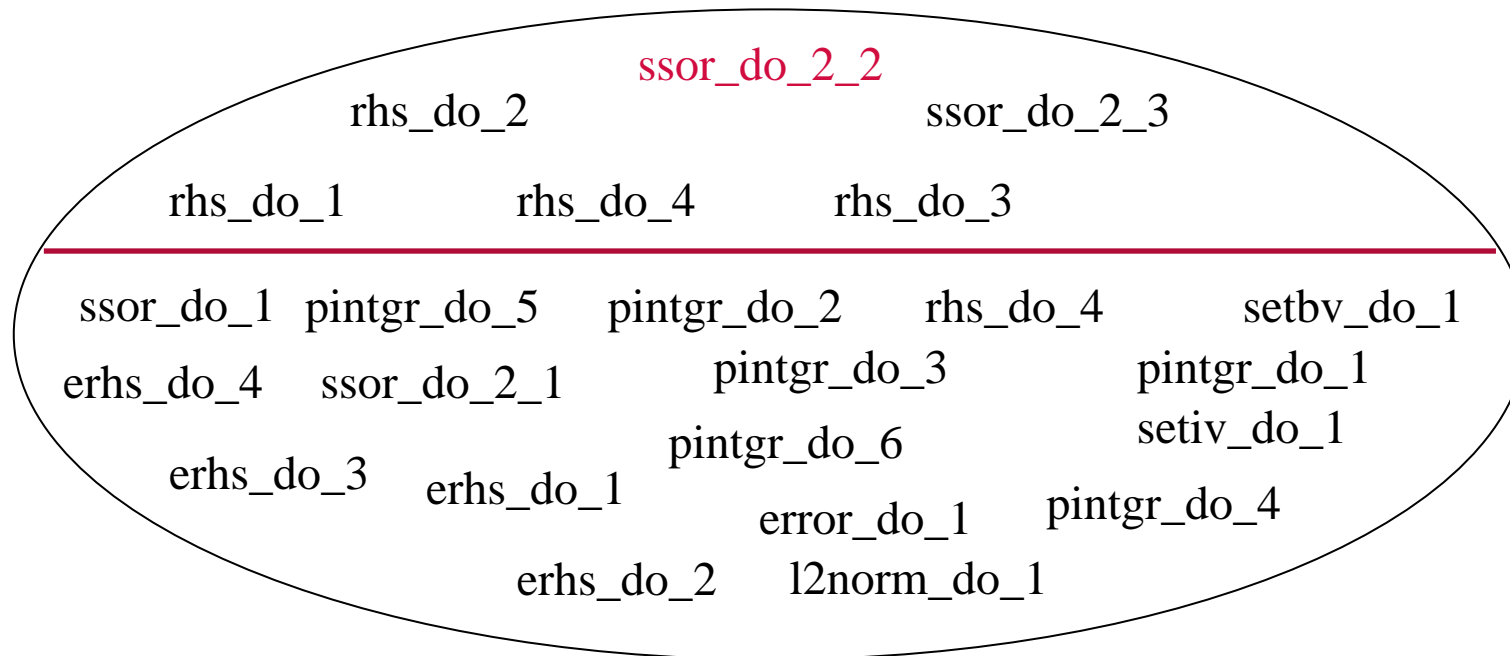
Overall Performance of APPLU



Overall Performance of APPLU

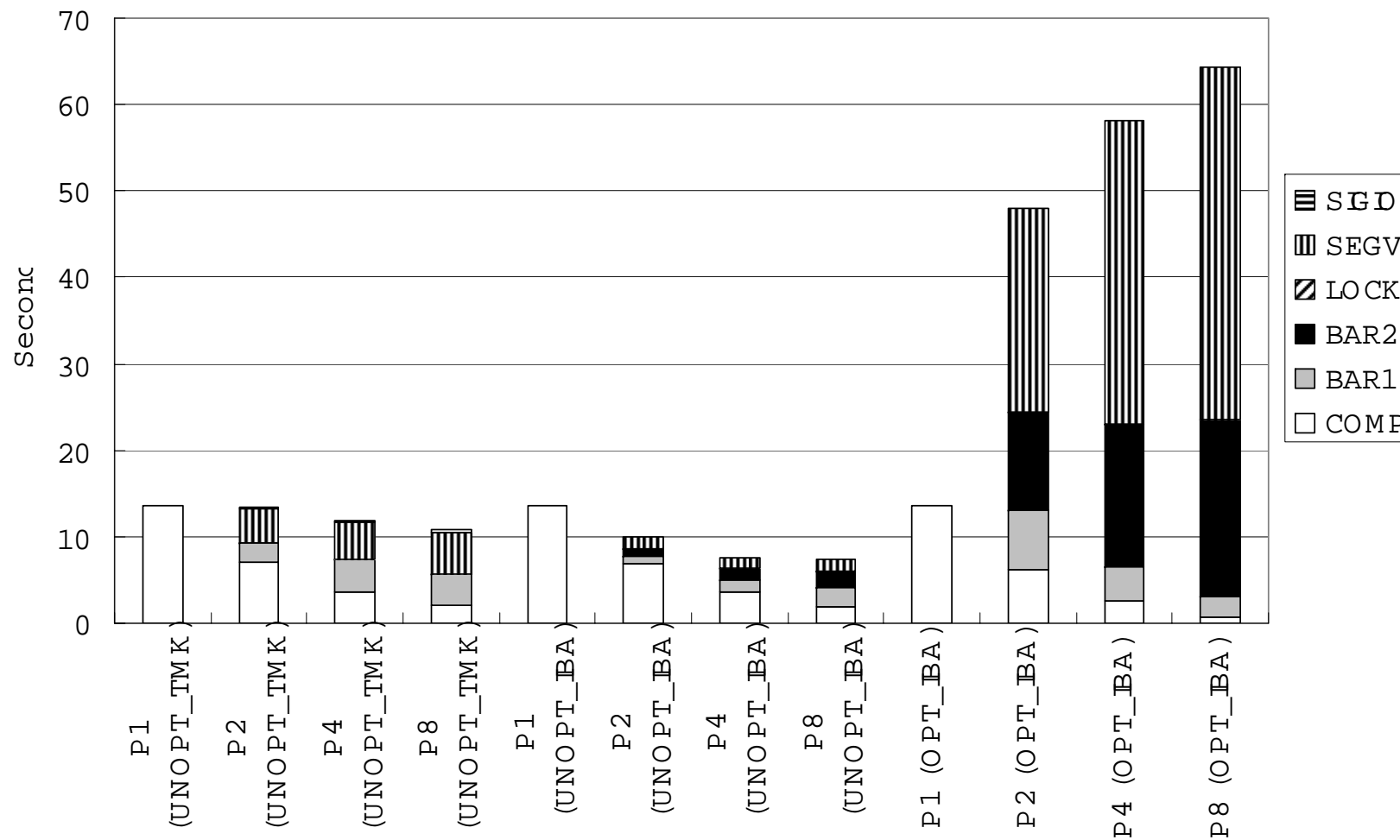


Parallel Loop



Serial Loop

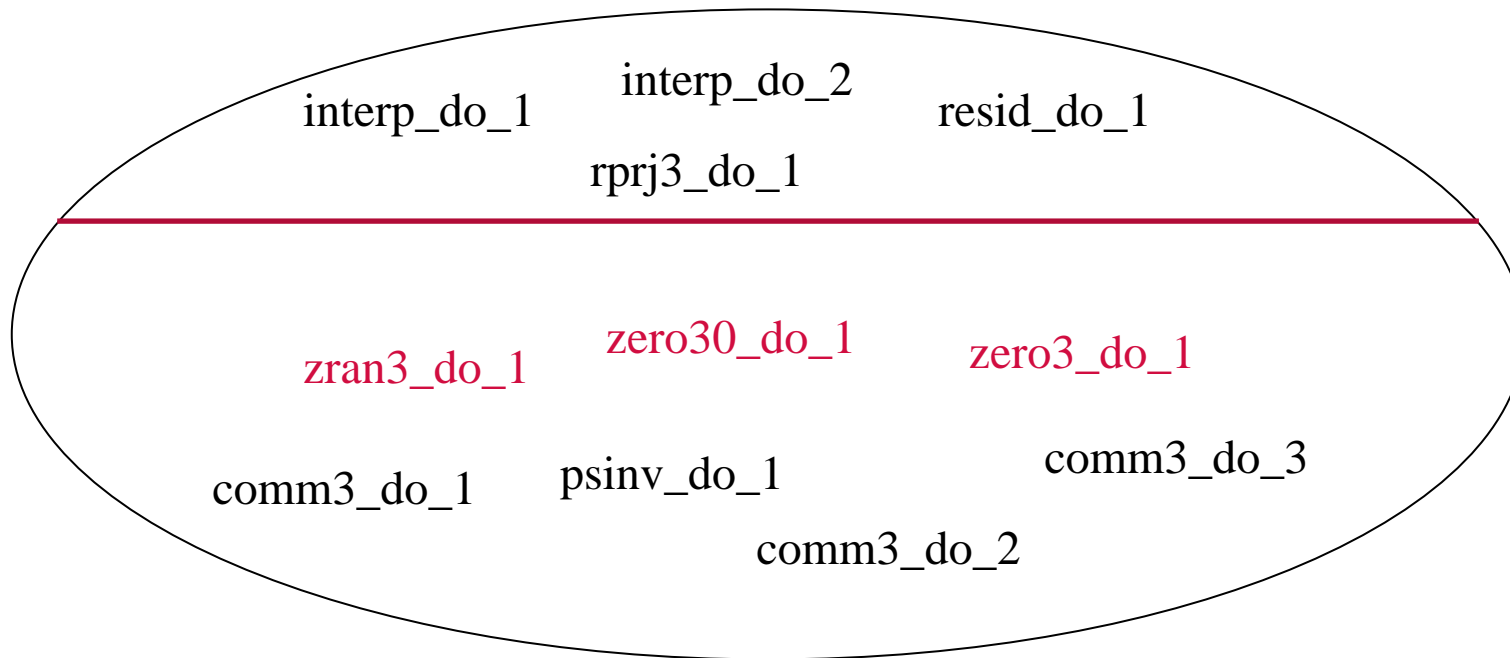
Overall Performance of MGRID



Overall Performance of MGRID

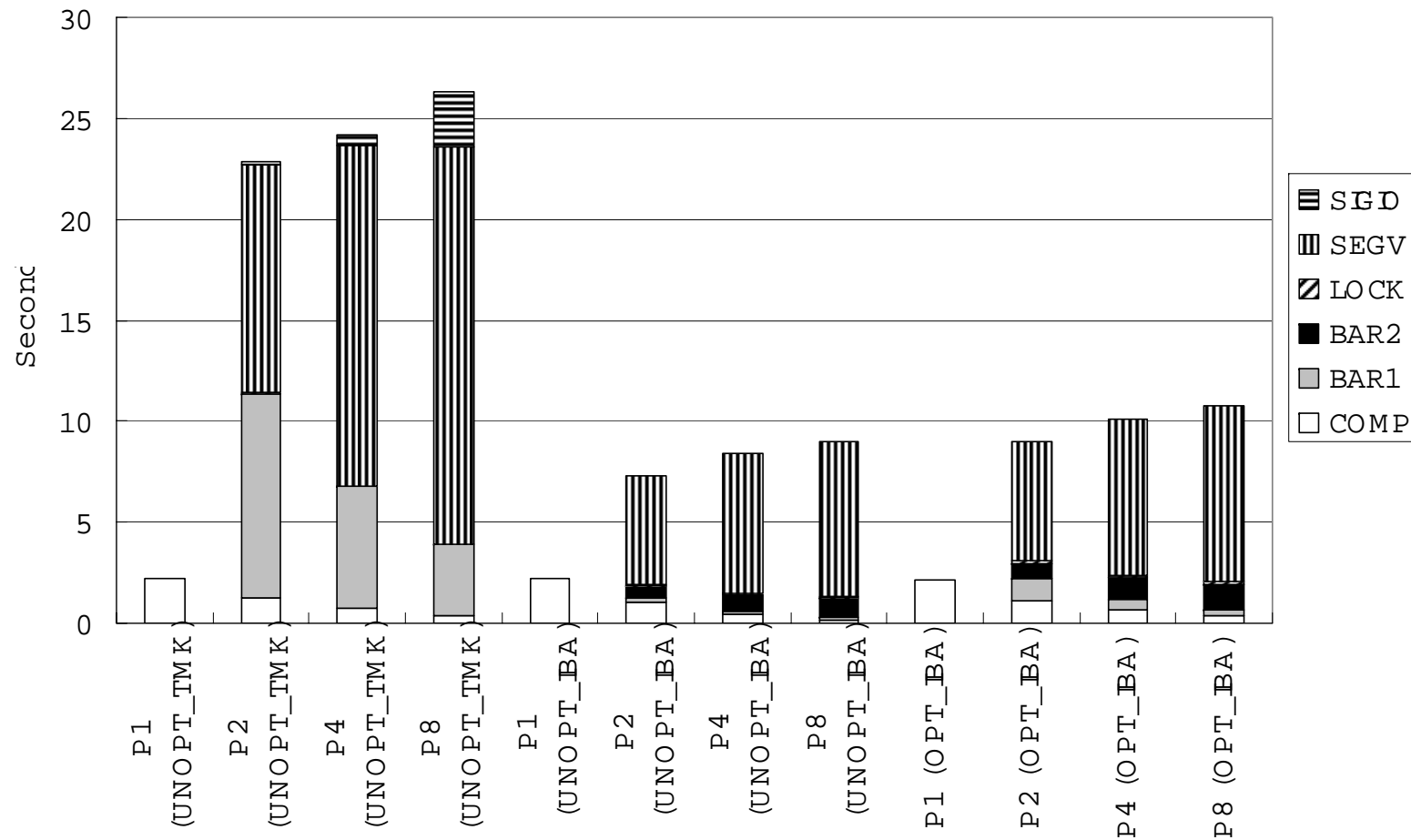


Parallel Loop



Serial Loop

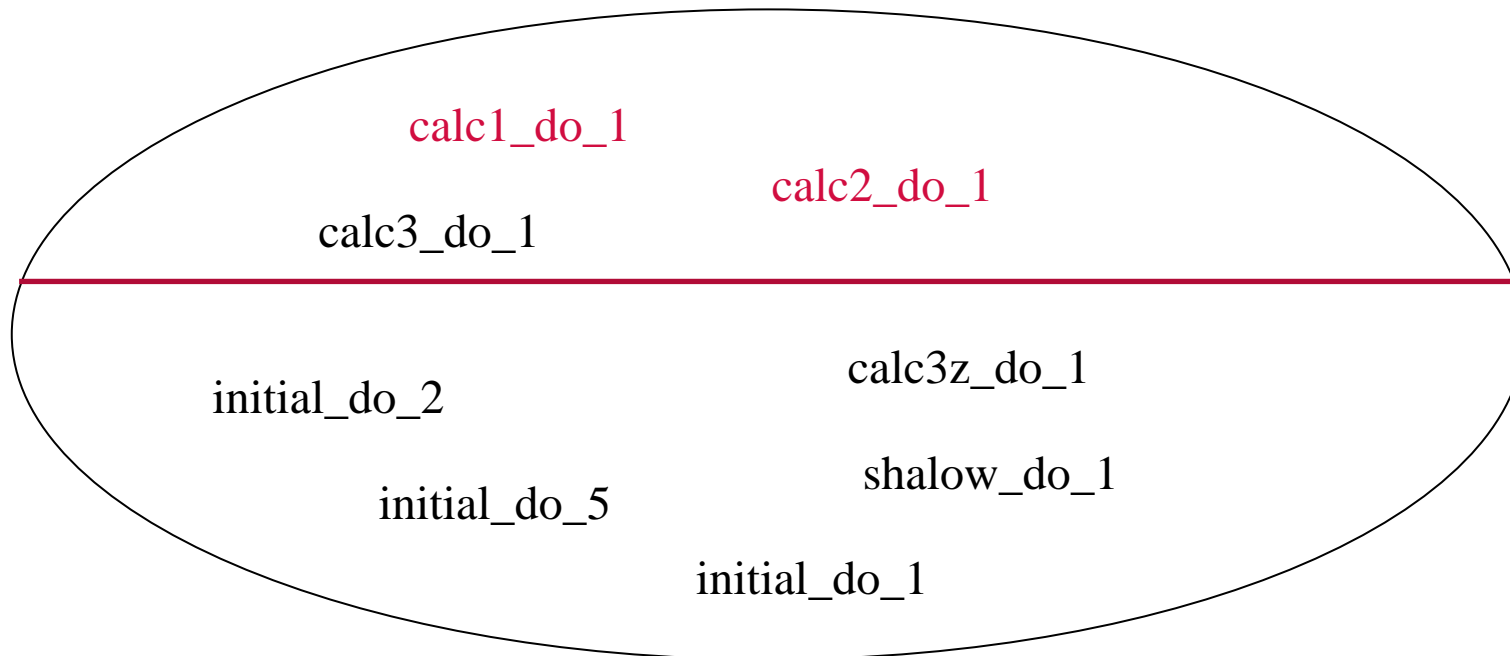
Overall Performance of SWIM



Overall Performance of SWIM

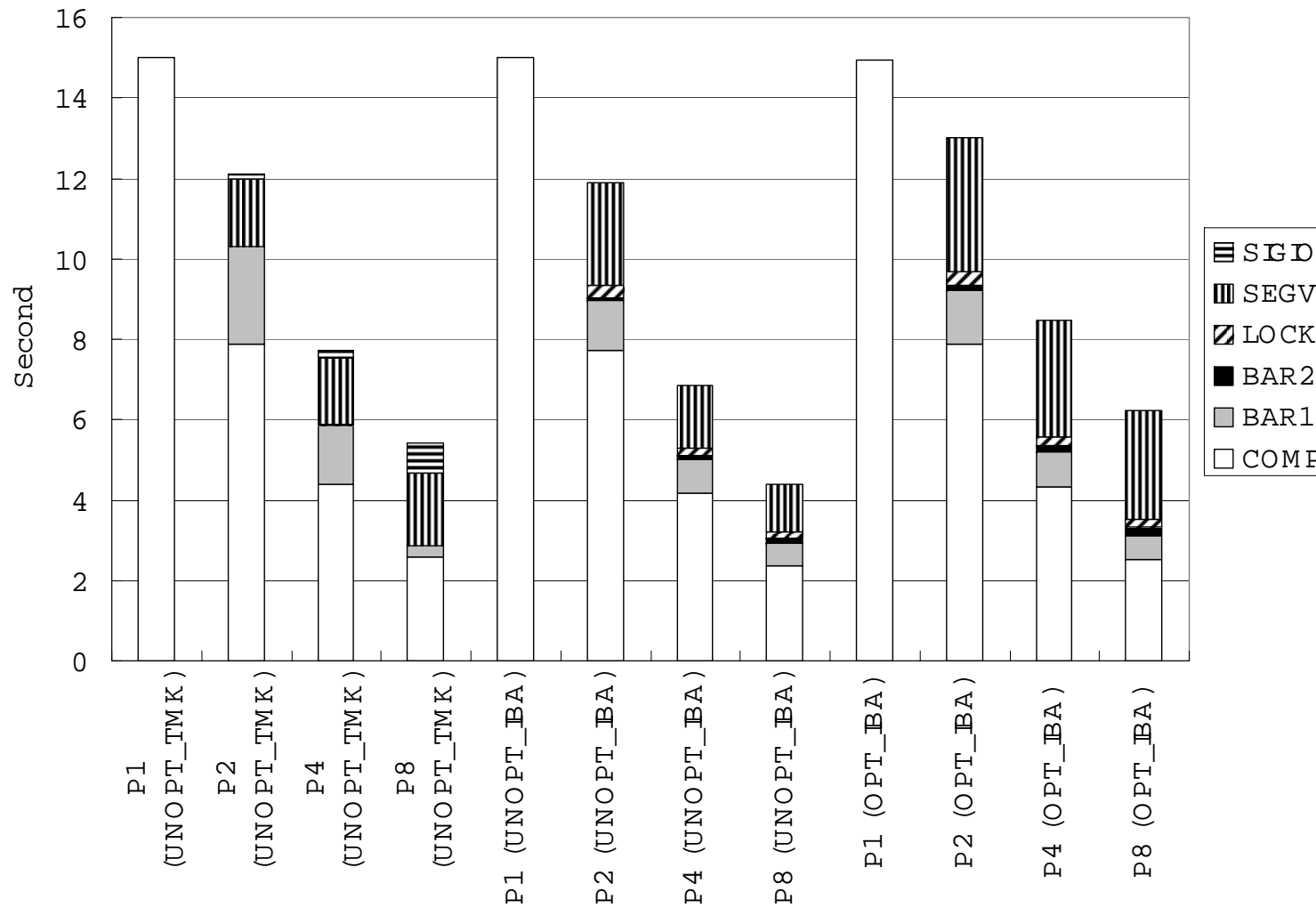


Parallel Loop



Serial Loop

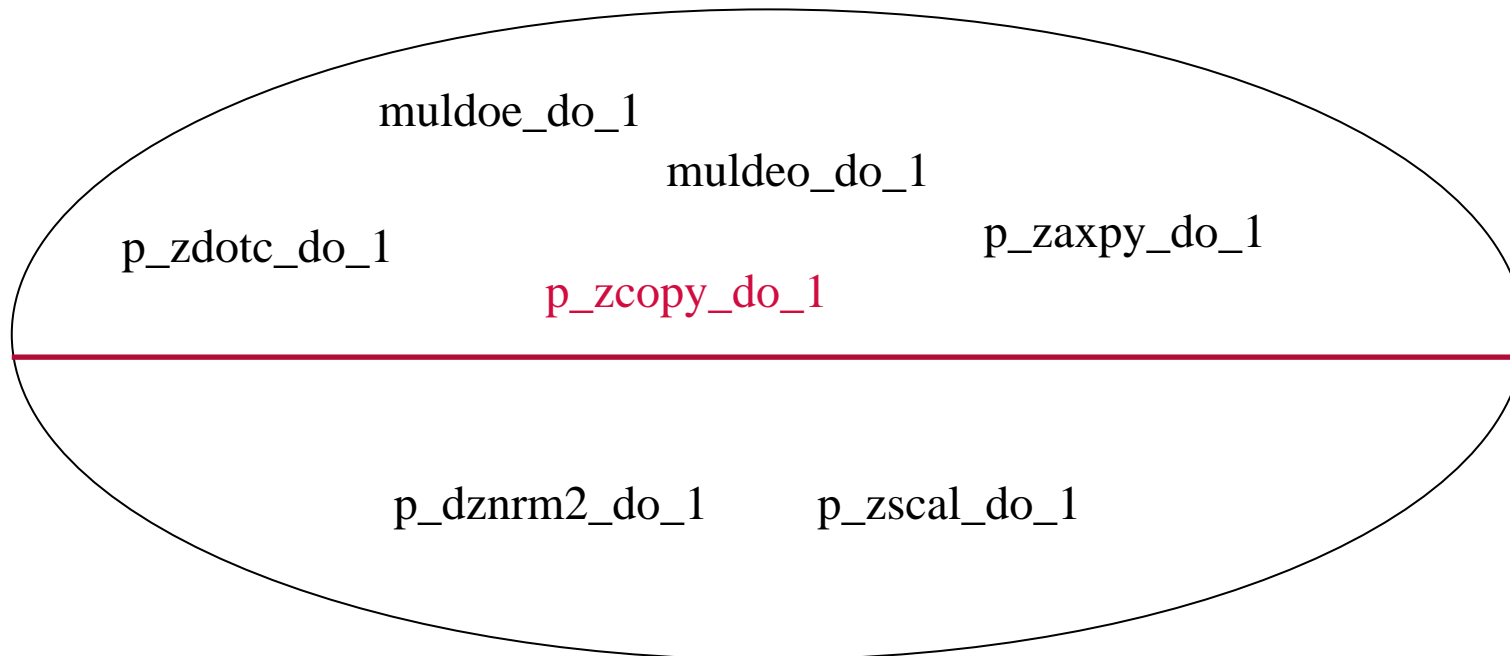
Overall Performance of WUPWISE



Overall Performance of WUPWISE



Parallel Loop



Serial Loop

Conclusion



- We showed that the next-generation of network technologies greatly improve the performance over the traditional implementation of DVSM systems.
- The performance gain results from using the remote DMA features, which does not involve any process interruption for data communication.
- Also, the serialization of parallel code sections to reduce parallel code overhead results in severe overhead on the DVSM systems due to more data movement between serial and parallel sections.
- Application performance is very sensitive to the data locality

Thanks

- Advanced Computer Systems and Compiler Lab:
 - <http://compiler.korea.ac.kr>