Server Oriented Microprocessor Optimizations

Charles R. Moore
Senior Technical Staff Member
crmoore@us.ibm.com
IBM Corporation
What is a Server?

- Many different types of servers in use today (many more tomorrow)
- All have interesting technical challenges and business opportunities
- The architecture of this collection of servers is a very interesting topic
- Today, I am focusing mostly on the Enterprise Server
Elements of Enterprise Server Performance

- Large system parallelism and concurrent execution
  - Tightly-coupled SMP scaling
  - NUMA access ratios
  - Clustering topologies

- Memory and I/O system design
  - Cache structure, Coherency protocols, "Smart" caching
  - Latency and Bandwidth
  - Network and I/O "impedance matching"

- Software optimization and path length
  - OS, Database, Application - algorithms and scaling
  - Compiler exploitation of hardware resources

- Compatibility and upgradability
  - Hot plug I/O, Disks, Memory, and Processors
  - Compatibility and durability between generations of machines
  - Logical and physical partitioning (dynamic reconfiguration)

- Reliability, Availability and Serviceability (RAS)
Q: Which system has better performance?

For servers, this is proving to be more important than Raw Performance!
Server Workload Characteristics

- Commercial
  - Large database footprints
  - Small record access
  - Random access patterns
  - Sharing/Thread communication

- Technical
  - Structured data
  - Large data movement
  - Predictable strides
  - Minimal data reuse

e-Business applications include attributes from both Commercial and Technical workloads
Today, processors spend most of their time waiting for cache misses. This is true for most workloads regardless of processor architecture or design. Feeding processors is the principal performance challenge.

The memory hierarchy bottleneck will get worse over time. Processor speed will continue to improve faster than memory and cache speeds. Software design trends (object oriented programming, just-in-time compilation, etc.) will place increased load on the memory hierarchy. SMP and NUMA designs expand the problem.

Memory hierarchy bandwidth and latency are limiting factors around which server designs need to be optimized.
Examples of Cache / Memory System Optimizations

1. Improve cache performance
   - on-chip cache hierarchy
   - exploitation of eDRAM technology for large caches
   - "smart caches" / adaptive cache coherency protocols
   - multiported caches and banking schemes
   - software controls for caches and TLBs
     (hints, prefetch, blocking, affinity, etc)

2. Manage overall latency
   - OOO execution to accelerate storage access instructions
   - multiple outstanding cache misses
   - hardware initiated prefetching (data and instructions)
   - allow speculation beyond synchronization boundaries
   - allow speculation beyond lock structures
Examples of Cache/Memory System Optimizations

(continued)

3. Maximize bandwidth
   - exploit extraordinary amount of available on-chip bandwidth
   - exploit large number of available module I/Os (cost trade-off)
   - fast I/O circuits and smart interface protocols

4. Multiprocessor optimizations
   - shared caches
   - efficient cache invalidate (XI) and cache-to-cache transfers
   - minimize synchronization / barrier overhead (avoid broadcasts)
   - fast lock processing; dedicated lock fabric between processors
   - Exploit weak storage consistency model (posted stores)
   - Multiple Threads per Chip (CMP, HMT, SMT)
Technology Effects on SMP Performance

Scattered Technology Deployment
- Curve flattens out quickly
- Inherent limitations work against you

Synergistic Technology Deployment
- Better scaling ratios
- More usable processors
- Higher overall throughput

SMP performance strongly benefits from synergistic technology deployment
Potential Architecture Optimizations for Servers

- Synchronization, Locking, and Cache Controls
  - Special purpose synchronization ops - only pay for what you need
  - Dedicated lock hardware
  - Cache policy hints

- Special Purpose accelerators
  - Move, Copy, Zero, Compare pages
  - Pointer chasing acceleration
  - Programmable stream prefetching engine

- Error recovery and RAS
  - Synchronous machine checks on memory / bus errors
  - Multiple interrupt tolerance

- Support for NUMA and Clustering
  - Message passing optimizations; Broadcast optimizations
  - Synchronous fencing of store errors

- Support for Logical Partitioning

In Servers, the ISA is far less important than the system-level optimizations.
Attributes of Server Oriented Microprocessors

- Choppy workloads; modest amounts of ILP
- Workloads have large instruction and data footprints
- Workloads demonstrate high degree of data sharing
- Workload partitioning ranges from trivial to very complex
- Complex, multi-tiered SW and system environments
- Systems demand non-stop operation (e-business)
- Systems demand configuration flexibility

High Frequency Operation

Optimized memory systems with large caches

Shared caches; Optimized intervention

Optimized Locking and Synchronization

Support tight SMP, NUMA & Clustering

Full system design and optimization

Strong focus on RAS

Binary compatibility across generations

Architecture extensions for partitioning
IBM's GigaProcessor (POWER4)

- Cornerstone of significant new Enterprise System Architecture
  - RS/6000 and AS/400 Systems
  - Binary compatibility with previous systems
  - Enhancements for synch, locking, partitioning, compiler controls

- > 1 GHz Operating Frequency (starting point)
  - Full custom design leveraging copper wiring and SOI

- Dual processors, integrated L2 Cache and L3 Cntrl on CPU chip

- Aggressive, SMP optimized Cache Hierarchy
  - Low latency access, very high bandwidth
  - High bandwidth cache-to-cache interconnection fabric
  - Hardware-based prefetching for instructions and data

- Enterprise-class RAS features

- Development substantially far along
POWER4 - Chip Multiprocessing

>1 Ghz Core  >1 Ghz Core

Shared L2

>100 GB/s Bandwidth
POWER4 - High BW L3 and Memory

>1 Ghz Core >1 Ghz Core

Shared L2

L3 Dir

L3

Memory

>10 GB/s Bandwidth
POWER4 - Low-end Server Solution

>1 Ghz Core

>1 Ghz Core

Shared L2

L3 Dir

L3

Memory

Expansion Bus

Expansion
Server Building Block

- >1 Ghz Core
- >1 Ghz Core
- Shared L2
- L3 Dir
- Chip-chip communication
- L3
- Memory

>35 GB/s Chip Interconnect

>100 GB/sec L2 to Core BW
>10 GB/sec L3 BW
>35 GB/sec Interconnect BW
Server Multi-chip Module (8-way SMP)

- Memory
- Expansion Bus
- Memory
- Expansion Bus
- Chip-chip communication
- Shared L2
- >1 Ghz Core
- L3 Dir.
- L3 Memory
- Expansion Bus
- Memory
POWER4 Unit-level Floorplan
~2300 Signal C4s

> 500 MHz
Wavepipelined I/O

> 1 Terabit/sec
Bandwidth at the Chip
POWER4 Multi-Chip Module
GigaProcessor Test Chip Die Photo
Technology Leverage in POWER4

- **Process**
  - IBM CMOS 8S2, 0.18um
  - Copper and SOI with 7 layers of metal
  - 170 million transistors

- **Package**
  - Uses large number of I/Os at chip and MCM level
    - >2,300 I/O with >5,500 Pins
  - Multi Chip Module (MCM) for dense integration

- **High bandwidth with fast busses**
  - Elastic I/O provides >500 Mhz chip-to-chip busses