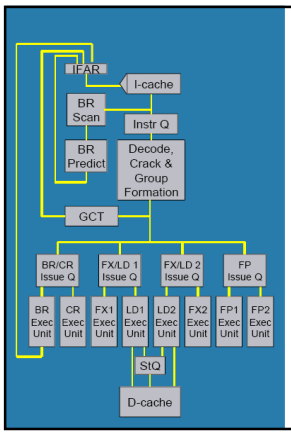
IBM eServer: (Discovery)

Power4 System Micro-architecture

A Summary

By Thomas Thomas



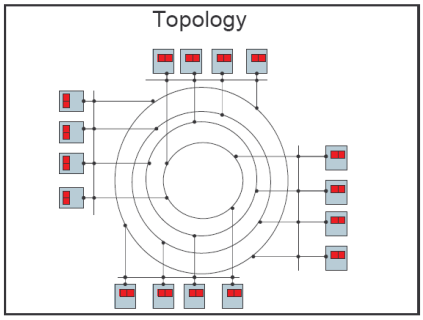
- Paper goes into a lot of detail about the core.
- Branch Prediction is paramount to long pipelines.
- Aggressive instruction fetching and pre-caching.
- Groups instructions into groups to ease logic of branch/cache misses.
- Also, marketing about 32/64 bit support, iSeries, pSeries, etc. not real interesting for us.
- 252, 254 stuff.

Multi-processor Organization

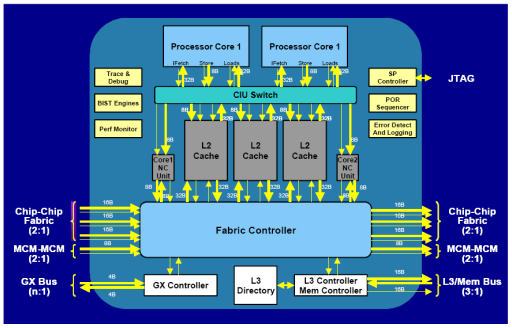
- 2 processing cores per chip.
- 4 chips per module
- 1, 2, 3, or 4 modules per SMP.
- That's up to 32 hardware contexts running concurrently
- Each chip has it's own cache and cache controllers for system balance.

Processor Communication

- Logically there are four, 16-byte buses for on module communication.
  - Implemented with 6 buses: 3 on and 3 off
- There are two, 8-byte buses for inter-module communication.
  - 1 on and 1 off
- Other buses for L3, IO devices, and system service processor. (GX bus)
- All processors use their own bus, and snoop everyone else's.



Chip Organization



Out of Order Load/Store Unit and Cache Coherence

"Decode, Crack and Group Formation:
As instructions are executed out of order, it is necessary to remember the program order of all instructions in flight. In order to minimize the logic necessary to track a large number of in flight instructions, groups of instructions are formed. The individual groups are tracked through the system. That is, the state of the machine is preserved at group boundaries, not at an instruction boundary within a group. Any exception causes the machine to be restored to the state of the oldest group prior to the exception."

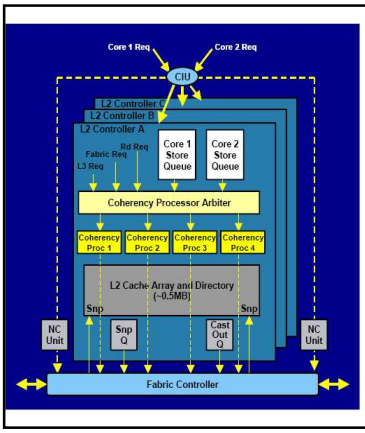
Continued

- Even with instructions in groups there is a lot of book keeping, with "mundane" instructions, others need special attention.
- Load, Hit, Store: Load executing before older store finishes writing. (forwarded data)
- Store, Hit, Load: Load executing before we realize there is an older store. (stale data)
- Load, Hit, Load: Old/new load consistency.

Memory Hierarchy and coherency

Component	Organization	Capacity per Chip
L1 Instruction Cache	Direct map, 128-byte line managed as 4 32-byte sectors	128 KB (64 KB per processor)
L1 Data Cache	2-way, 128-byte line	64 KB (32 KB per processor)
L2	8-way, 128-byte line	~ 1.5 MB
L3	8-way, 512-byte line managed as 4 128-byte sectors	32 MB
Memory	---	0-16 GB

The coherency point for POWER4 is the L2 cache. All data in the L1 data cache is also in the L2 cache.



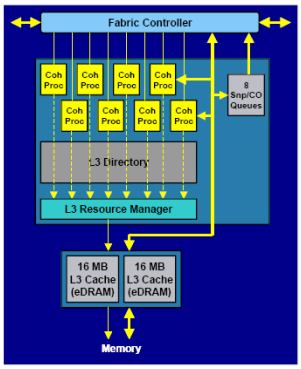
The L2 Cache

The coherency processor

- Controls the return of data from the L2 (hit) or from the fabric controller (miss) to the requesting core via the CIU;
- Updates the L2 directory as needed;
- Issues fabric commands for L2 misses on fetch requests and for stores that do not hit in the L2 in the M, Me or Mu state (described below);
- Controls writing into the L2 when either reloading due to fetch misses in the L2 or stores from the processors; and,
- Initiates back invalidates to a processor via the CIU resulting from a store from one core that hits a cache line marked as resident in the other processor's L1 data cache.

Coherency States (different from Sun's MESI)

- **I (invalid state): Invalid.**
- **SL (shared state, can be source to local requesters):**
- **S (shared state):**
  - Somebody else will share
- **M (modified state):**
  - Exclusively owned, but can be sourced to others.
- **Me (exclusive state):**
  - Not considered modified, but I have requested exclusivity.
- **Mu (unsolicited modified state):**
  - Modified, and still exclusive.
- **T (tagged state):**
  - Was modified and sourced to another L2.



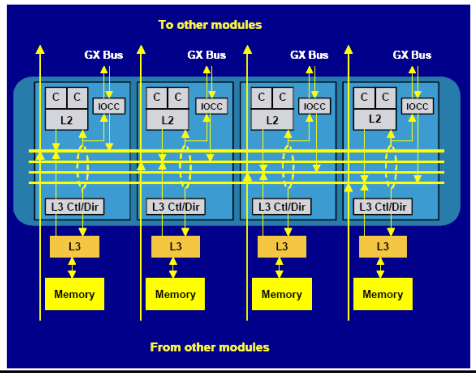
The L3 Cache.

The L3 is designed to be used as a standalone 32 MB L3 cache, or to be combined with other L3s on the same processor module in pairs or groups of four to create a larger, address interleaved L3 cache of 64 MB or 128 MB. Combining L3s into groups not only increases the L3 cache size, but also scales the available L3 bandwidth

L3 Continued

- **I (invalid state):** The data is invalid.
- **S (shared state):** The data is valid. In this state, the L3 can only source data to L2s
- that it is caching data for.
- **T (tagged state):** The data is valid. The data is modified relative to the copy stored in memory. The data may be shared in other L2 or L3 caches.
- **Trem (remote tagged state):** This is the same as the T state, but the data was sourced from memory attached to another chip.
- **O (prefetch data state):** The data in the L3 is identical to the data in memory. The data was sourced from memory attached to this L3. The status of the data in other L2 or L3 caches is unknown.

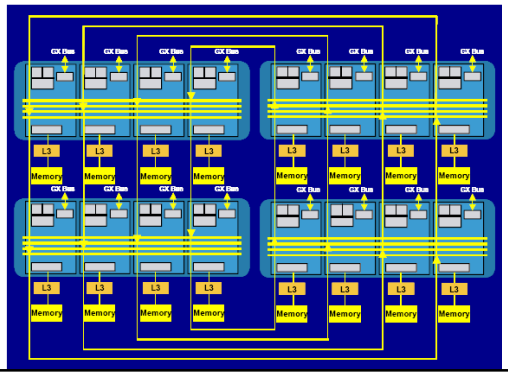
The Module



Multiple Modules

Multiple Module Interconnect:
Figure 9 shows the interconnection of multiple 4-chip MCMs to form larger SMPs. From 1 to 4 MCMs can be interconnected. When interconnecting multiple MCMs, the intermodule buses act as repeaters moving requests and responses from one module to another module in a ring topology. As with the single MCM configuration, each chip always sends requests/commands and data on its own bus but snoops all buses.

Multiple Modules, continued.



Power5

- 184 million transistors to 276 million transistors
- Increased cache size and bandwidth
- L3 now on processor side of fabric controller
  - And off the inter-chip bus
- Simultaneous multi-threading per core.
  - 2 threads.
- Up to 8 modules, or 128 active threads.
  - 4 books * 2 MCMs * 4 Processors * 2 cores * 2 threads