IBM POWER9

Bhopesh Bassi, Ivan Chen, Wes Darvin

What is POWER9

- IBM's POWER processor line
- Servers and high-compute workloads
  - Analytics, AI, cognitive computing
  - Technical and high-performance computing
  - Cloud/hyperscale data centers
  - Enterprise computing
- Summit Supercomputer @ Oak Ridge National Lab
  - 200 petaflops

[1,7]

Multithreading and Multiprocessing

Multithreading and Variants

- 12 core and 24 core variants
  - 12 x SMT8 cores
  - 24 x SMT4 cores
- SMT8 supports simultaneous multithreading of up to 8 threads
- SMT4 supports up to 4
- Total resources the same, divided differently

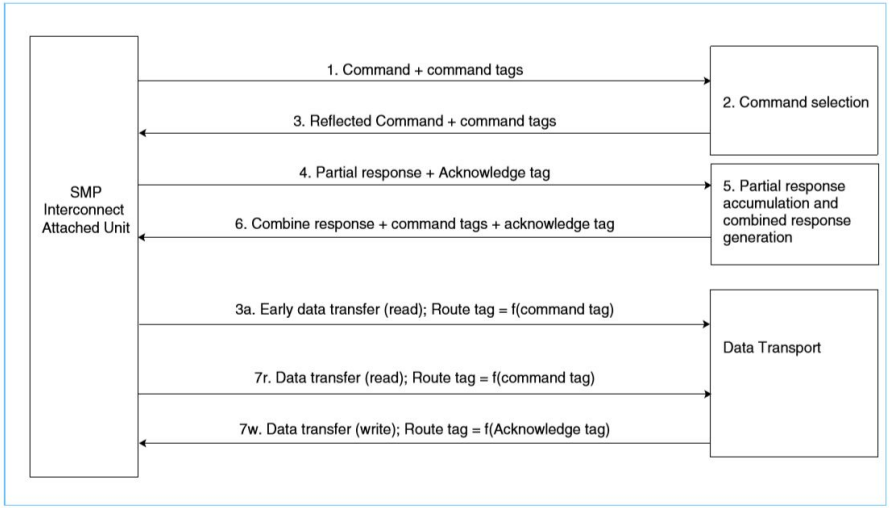
- SMT8 is optimized for IBM's PowerVM (server virtualization) ecosystem
- SMT4 is optimized for the Linux Ecosystem

[1]

Symmetric Multiprocessing Interconnect

- Hardware to enable cache-coherent communication between processors
- Two external SMP hookups to connect other POWER9 chips
- Snooping based protocol
  - Multiple command and response scopes to limit bandwidth use

Figure 8-1. SMP Interconnect Coherency Protocol



[2]

Core Microarchitecture

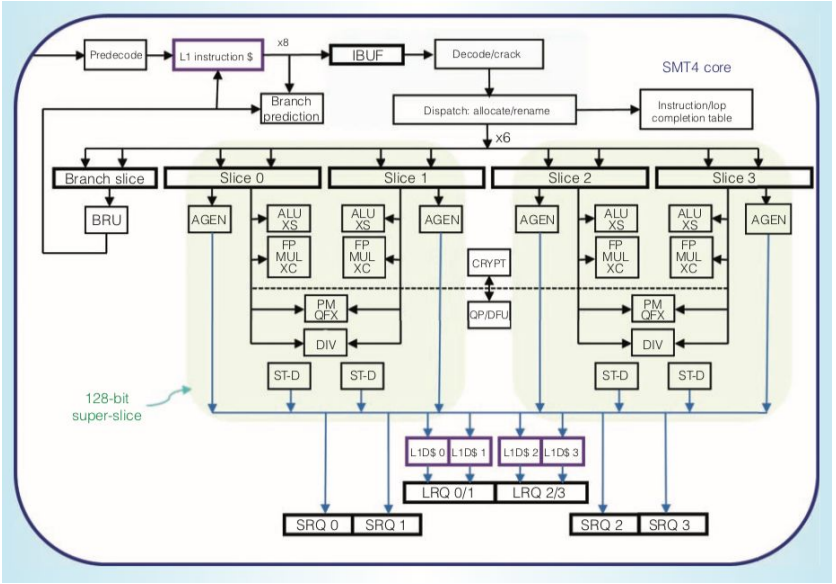
Pipeline Structure

- Single Front-End(Master) Pipeline
  - Allows for speculative in-order instructions
  - Throws away mispredicted paths
- Multiple executional unit pipelines
  - Allows for out-of-order instructions of both speculative and non-speculative operations
- Execution Slice Microarchitecture
- Pipeline supports completion of up to 128 instructions per cycle (SMT4 Core)
  - Completion of 256 instructions per cycle
- 32KB, 8-way assoc I-Cache and D-Cache
- One-cycle to preprocess inst.
  - Up to six instructions decoded concurrently

[1, 2,]

Slice Microarchitecture

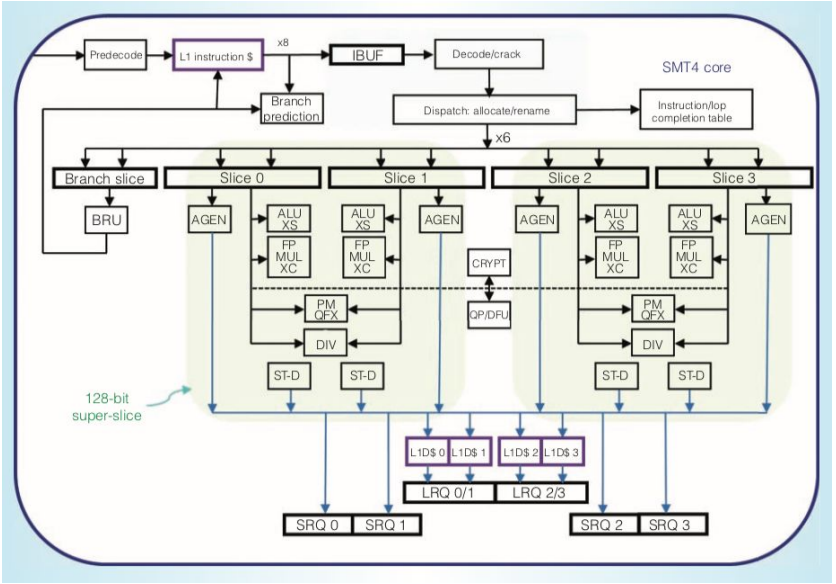
- 4 Executional slices and 1 Branch Slice
  - 2 execution slices form a super slice and 2 super slices combine to form a four-way simultaneous multithreading core(SMT4 Core)
- 128-entry Instruction Completion Table(SMT4 Core)
- History Buffer and Reorder Queue for out-of-order execution
  - Each of the 4 slices have a history buffer and reorder queue



[1, 2]

Slice Microarchitecture

- Four Fixed-Point and LD/ST execution pipelines; One FP Unit and Branch Execution pipeline
- Four Vector Scalar Units
  - Binary FP pipeline
  - Simple and Complex Fixed-Point pipeline
  - Crypto Pipeline
  - Permute Pipeline
  - Decimal floating point pipeline



[1, 2]

Branch Prediction

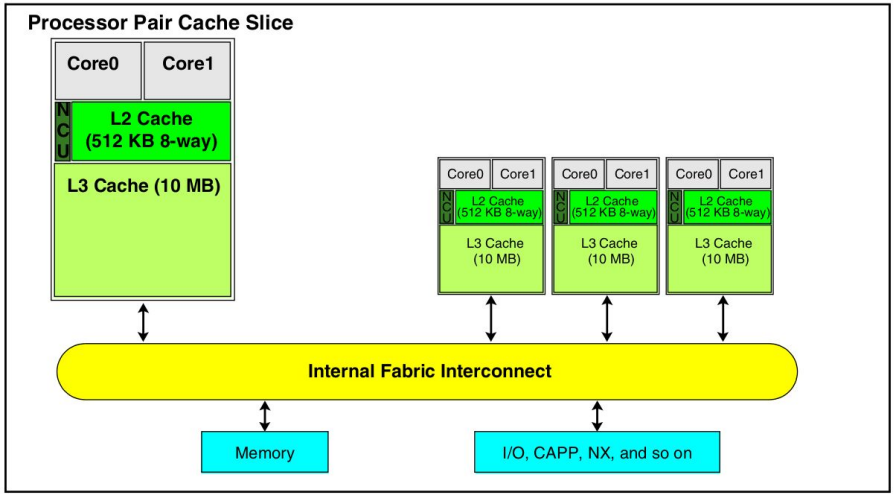
- Direction and Target Address Prediction
- Predict up to 8 branches per cycle
- Static and Dynamic Branch Prediction
  - Static Prediction based on Power ISA
- Four branch history tables: global predictor, local predictor, selector, local selector
  - Used for Dynamic Prediction
  - Each prediction table has 8K entries x 2bit
- Other methods:
  - Link Stack, Count Cache, Pattern Cache

[1, 2]

Cache and Memory subsystems

Cache Hierarchy Overview for SMT4 variant

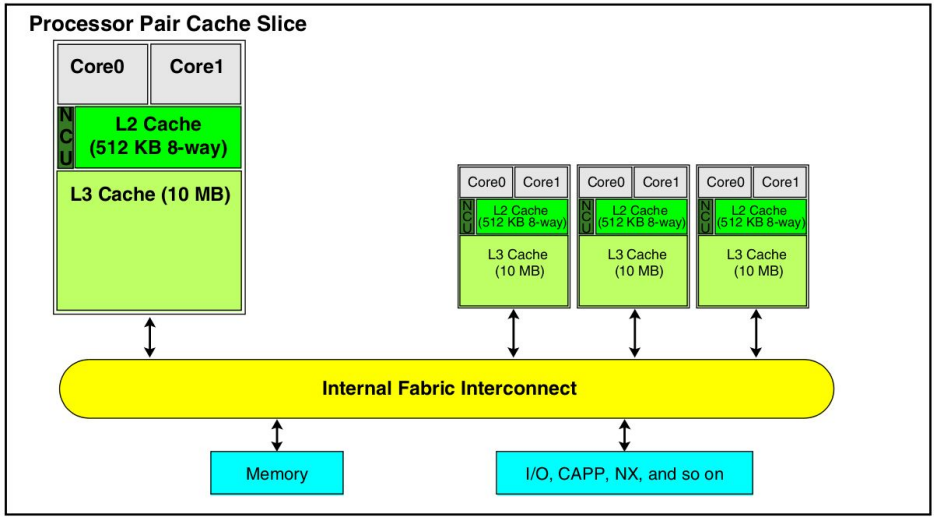
- Three level cache
- 128 byte cache line
- Physically indexed physically tagged
- L1:
  - Separate ICache and DCache
  - 32 KB 8 way
  - Store through and no write allocate
  - Pseudo LRU replacement
  - Includes way predictor



[1, 2, 4, 5]

Cache Hierarchy Overview for SMP4 variant contd..

- L2:
  - 512 KB 8-way Unified
  - Shared by two cores
  - Store back write allocate
  - Double banked
  - LRU
  - Coherent
- L2 is inclusive of L1
- L3:
  - 120 MB **shared by all cores**.
  - **Victim cache for L2** and other L3 regions
  - **NUCA**(Non uniform cache architecture)
  - Each 10 MB region is 20 way set associative
  - Sophisticated replacement policy based on historical access rates and data types.
  - Coherent



[1, 2, 4, 5]

Prefetching

- Prefetch engine tracks loads and stores addresses.
- Recognizes streams of sequentially increasing/decreasing accesses.
  - N-stride detection
- Every L1 D-cache miss is a candidate for new stream.
- Confirmed access in stream causes engine to bring one additional line into each of L1, L2 and L3 cache.
- Upto 8 streams in parallel.
- Software initiated prefetching
- Mitigates cache pollution and premature eviction
  - Lines brought into L3 are several lines ahead of those being brought into L1.

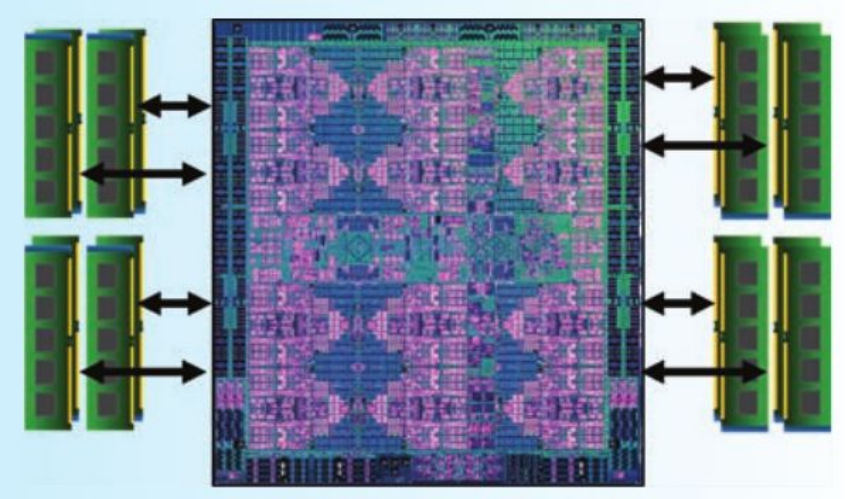
[2, 3]

Adaptive Prefetching

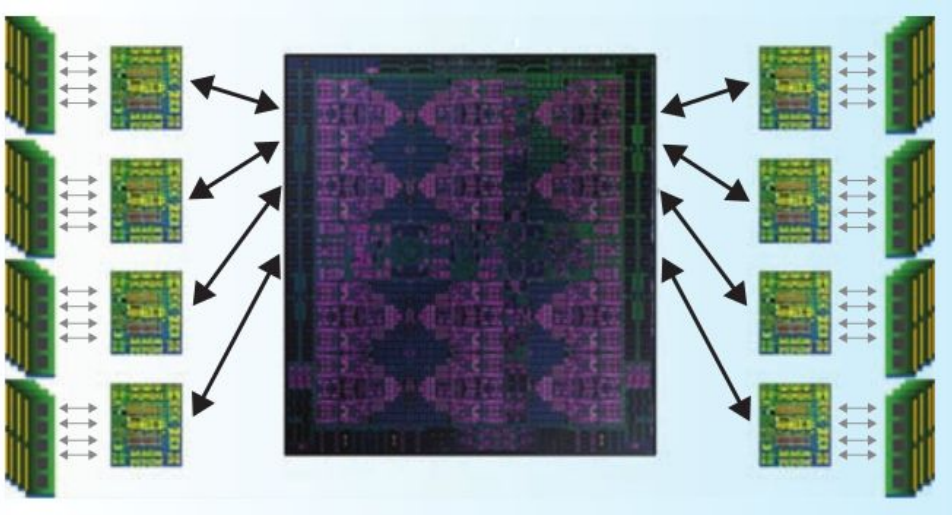
- Confidence levels associated with prefetch requests.
  - Determined based on program history and stream
- Memory controller prioritizes requests using confidence level
  - Crucial when memory bandwidth is low.
- Predicts phases of program where prefetching is more effective
- Receives feedback from memory controller to assist in determining depth of prefetch

[2, 3]

Memory subsystem



Directly attached memory
Scale-out version
Upto 4 TB



Agnostic Buffered memory
Scale-up version
Upto 8 TB

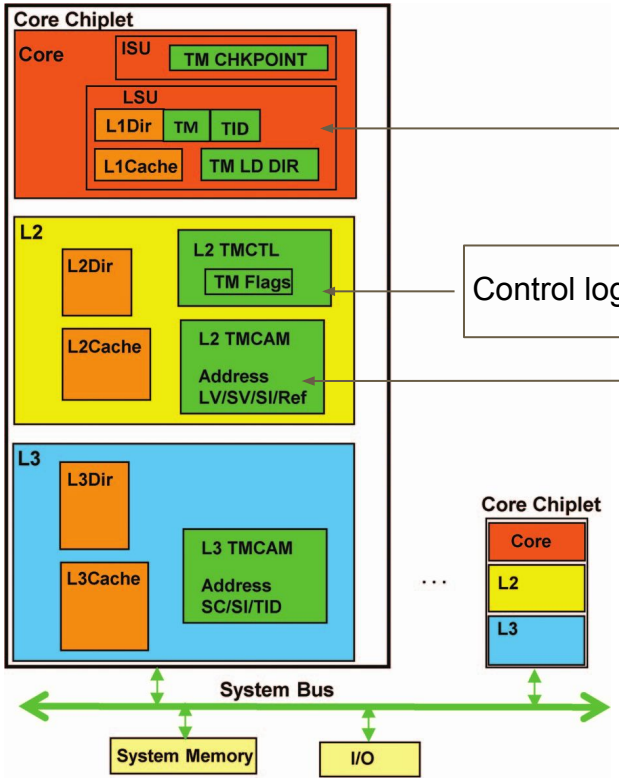
[1, 4, 5]

Transactions in Power8 and Power9

- Arbitrary number of loads and stores as a single atomic operation
- Optimistic concurrency control
- Better performance than locks when less contention
- Changes made by ongoing transaction not visible to other threads
- Possible conflicts:
  - Load-Store conflict between two transactions
  - Load-Store conflict between one transaction and one non-transactional operation.
- Implemented at hardware level in Power8 and Power9
  - ISA has instructions for starting, committing, aborting and suspending instructions
  - Best-effort implementation
  - Work with interrupts as transaction suspension is possible.

[6]

Transactions contd..



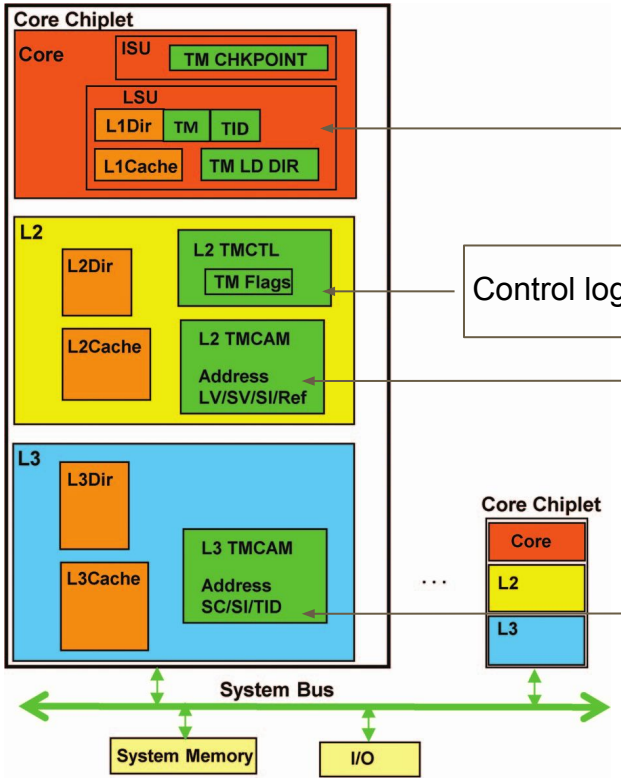
Control logic

L1 state per cache line
TM: Set if cache line part of store footprint of a transaction
TID: Thread id that did store to this cache line.

L2 state for each cache line
LV: Load valid. Set if cache line is in part of load footprint of one or more transactions
SV: Store valid. Set if cache is part of store footprint of a transaction
SI: Store Invalid. Set if transaction fails.
REF:
One bit per thread.
If LV is set, indicates which thread(s) are part of transactional load.
If SV is set, indicates which thread is part of transactional store.

[6]

Transactions contd..



Control logic

L1 state per cache line
TM: Set if cache line part of store footprint of a transaction
TID: Thread id that did store to this cache line.

L2 state for each cache line
LV: Load valid. Set if cache line is in part of load footprint of one or more transactions
SV: Store valid. Set if cache is part of store footprint of a transaction
SI: Store Invalid. Set if transaction fails.
REF:
One bit per thread.
If LV is set, indicates which thread(s) are part of transactional load.
If SV is set, indicates which thread is part of transactional store.

L3 state per cache line
SC: Set if cache line was dirty at the time of transactional store. Indicates that this is pre-transaction dirty copy of cache line.
SI: Set at transaction commit to indicate that pre-transaction copy is invalid.

[6]

Rollback only transactions

- Single thread speculative instruction execution
- Do not guarantee atomicity
  - Use only when accessed data is not shared with other threads
- Use case in trace scheduling
  - No need for complex compensation code.

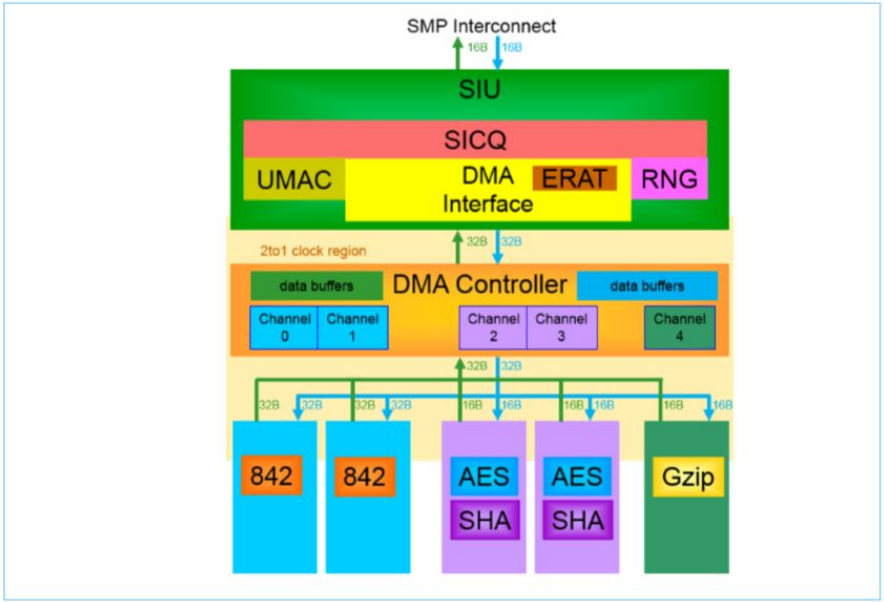
[6]

Heterogeneous Computing

On-chip Accelerators

- Nest Accelerator unit
  - DMA and SMP interconnect
  - 2 x 842 compression
  - 1 x GZip compression
  - 2 x AES/SHA

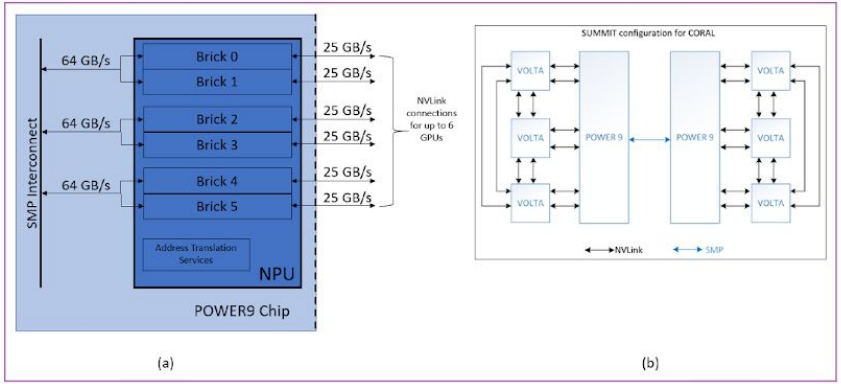
Figure 11-1. NX Block Diagram



[1,2]

GPUs / NVLink 2.0

- 25GB/s
  - 7-10x more bandwidth compared to PCIe Gen3
- Coherent memory sharing
- Access granularity
  - 1 - 256 bytes
- Flat address space
  - Automatic data management
  - Ability to manually manage data transfers



[1,8]

Coherent Accelerator Processor Interface

- POWER9 supports CAPI 2.0
- High bandwidth, low latency hookup for ASICs and FPGAs
- Allows cache coherent connection between attached functional unit to SMP interconnect bus

[1,2]

Questions

Sources

1. Power9 processor architecture: https://ieeexplore.ieee.org/document/7924241
2. Power9 user manual: https://ibm.ent.box.com/s/8uj02ysel62meji4voujw29wwkhsz6a4
3. Power9 core microarchitecture presentation: https://www.ibm.com/developerworks/community/wikis/form/anonymous/api/wiki/61ad9cf2-c6a3-4d2c-b779-61ff0266d32a/page/1cb956e8-4160-4bea-a956-e51490c2b920/attachment/5d3361eb-3008-4347-bf2f-6bf52e13f060/media/The%20Power8%20Core%20MicroArchitecture%20earlj%20V5.0%20Feb18-2016VUG2.pdf
4. Power9 memory: https://ieeexplore.ieee.org/document/8383687
5. Power8 cache and memory: https://ieeexplore.ieee.org/document/7029173
6. Power8 transactions: https://ieeexplore.ieee.org/document/7029245
7. ORNL Blogpost: https://www.ornl.gov/news/ornl-launches-summit-supercomputer
8. NVLink and POWER9: https://ieeexplore.ieee.org/document/8392669

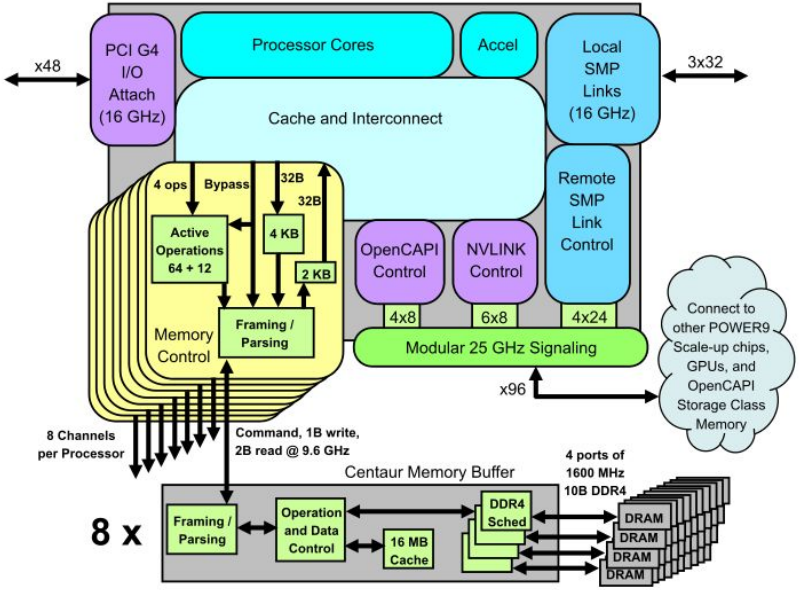
Backup Slides

SMP Interconnect

- Command broadcast scopes
  - Local Node Scope
    - Local chip with nodal (one chip) scope
  - Remote Node Scope
    - Local chip and targeted chip on a remote group
  - Group Scope
    - Local chip with access to the memory coherency directory
  - Vectored Group Scope
    - Local chip and targeted remote chip

DDR4 buffer chip: Centaur

- Centaur has 16 MB cache
- Acts as L4 cache
- Pros:
  - Lower write latency
  - Efficient memory scheduling
  - Prefetching extensions:
    - Prefetches prefetch requests for high confidence prefetch streams.
- Cons:
  - Load-to-use latency increases slightly
  - Complex system packaging



[4, 5]

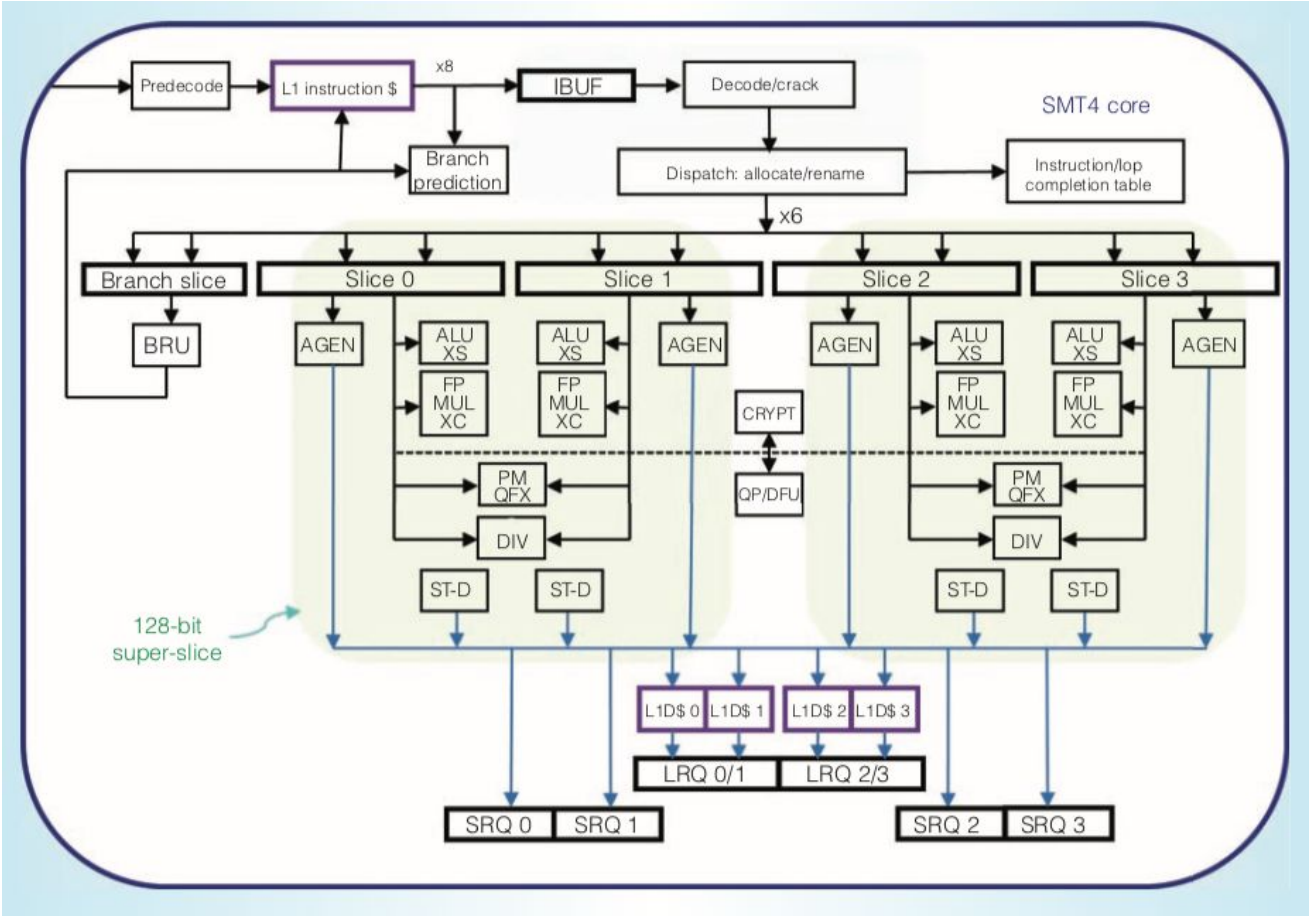


Diagram of slice microarchitecture