IBM POWER9

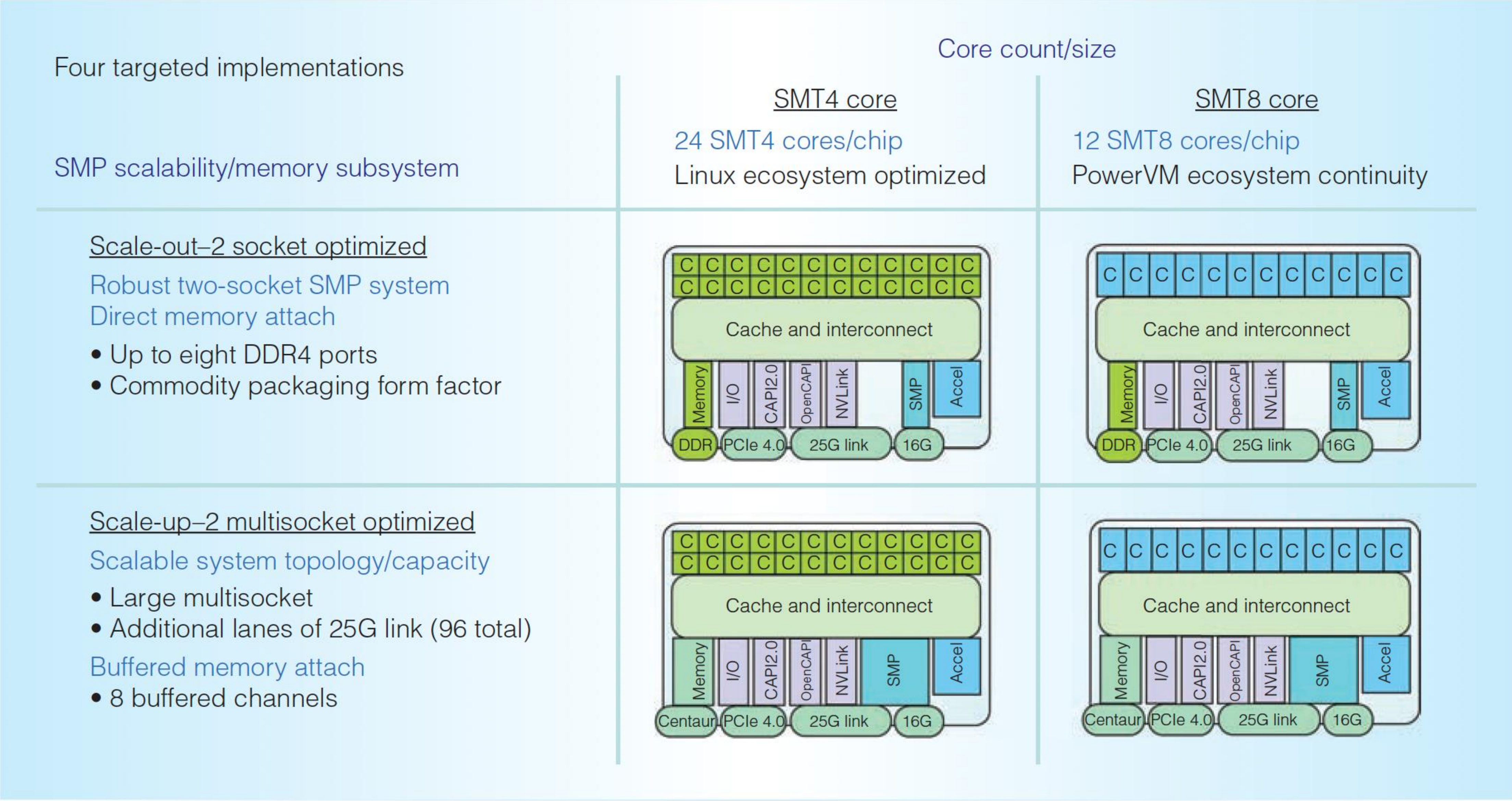
Talal Touseef, Talha Waheed, Muhammad Taimoor Tariq

{touseef2, twaheed2, mttariq2}@illinois.edu

Introduction

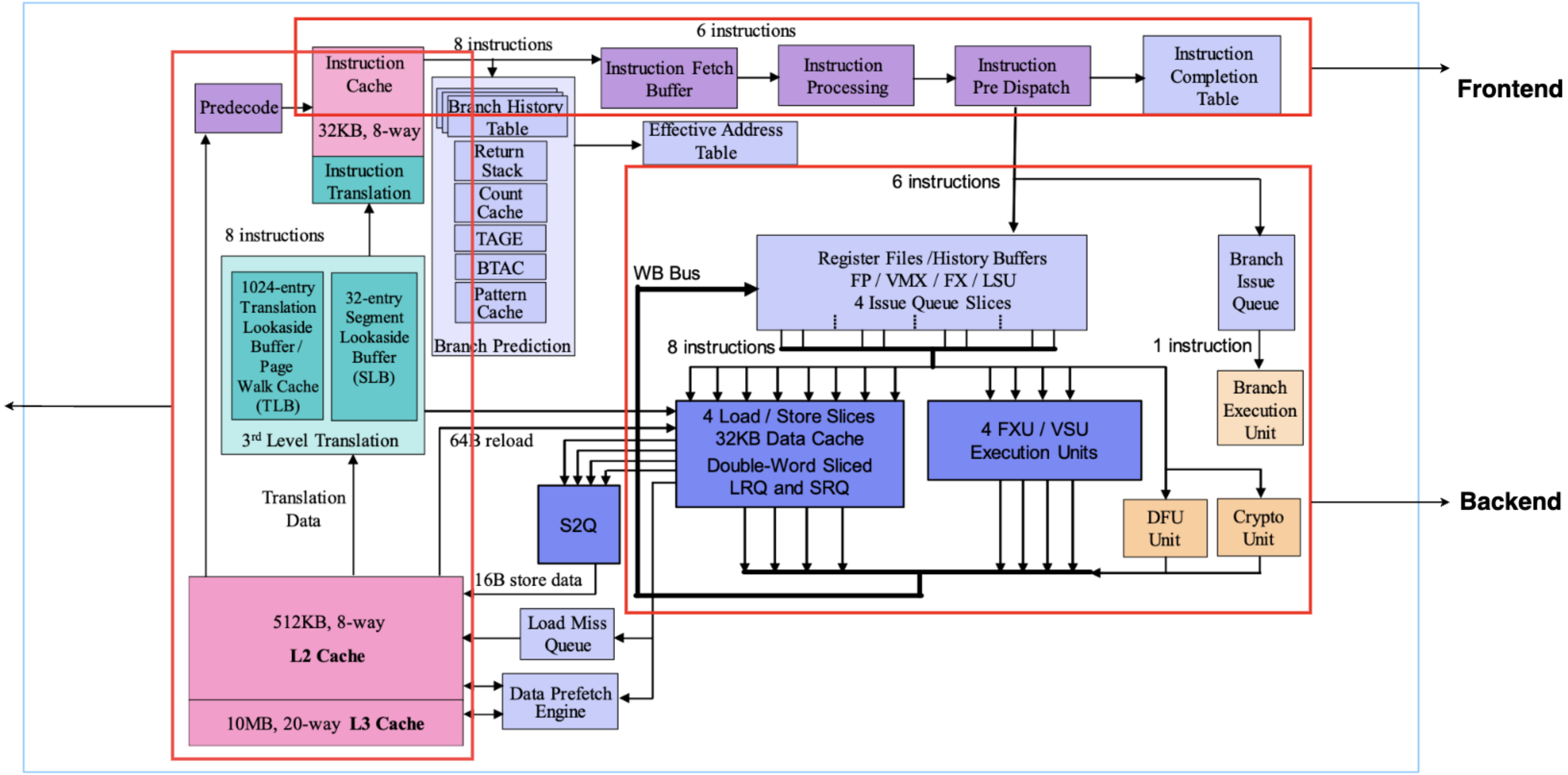
- Superscalar
- Symmetric (shared memory architecture) multiprocessor
- Designed for servers and large-cluster systems

Different target implementations

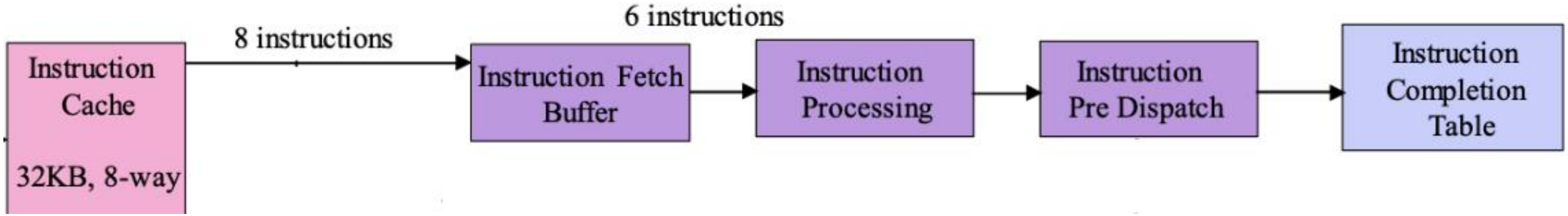


Four targeted implementations SMP scalability/memory subsystem	Core count/size SMT4 core 24 SMT4 cores/chip Linux ecosystem optimized	SMT8 core 12 SMT8 cores/chip PowerVM ecosystem continuity
Scale-out–2 socket optimized Robust two-socket SMP system Direct memory attach • Up to eight DDR4 ports • Commodity packaging form factor		
Scale-up–2 multisocket optimized Scalable system topology/capacity • Large multisocket • Additional lanes of 25G link (96 total) Buffered memory attach • 8 buffered channels		

Core Microarchitecture



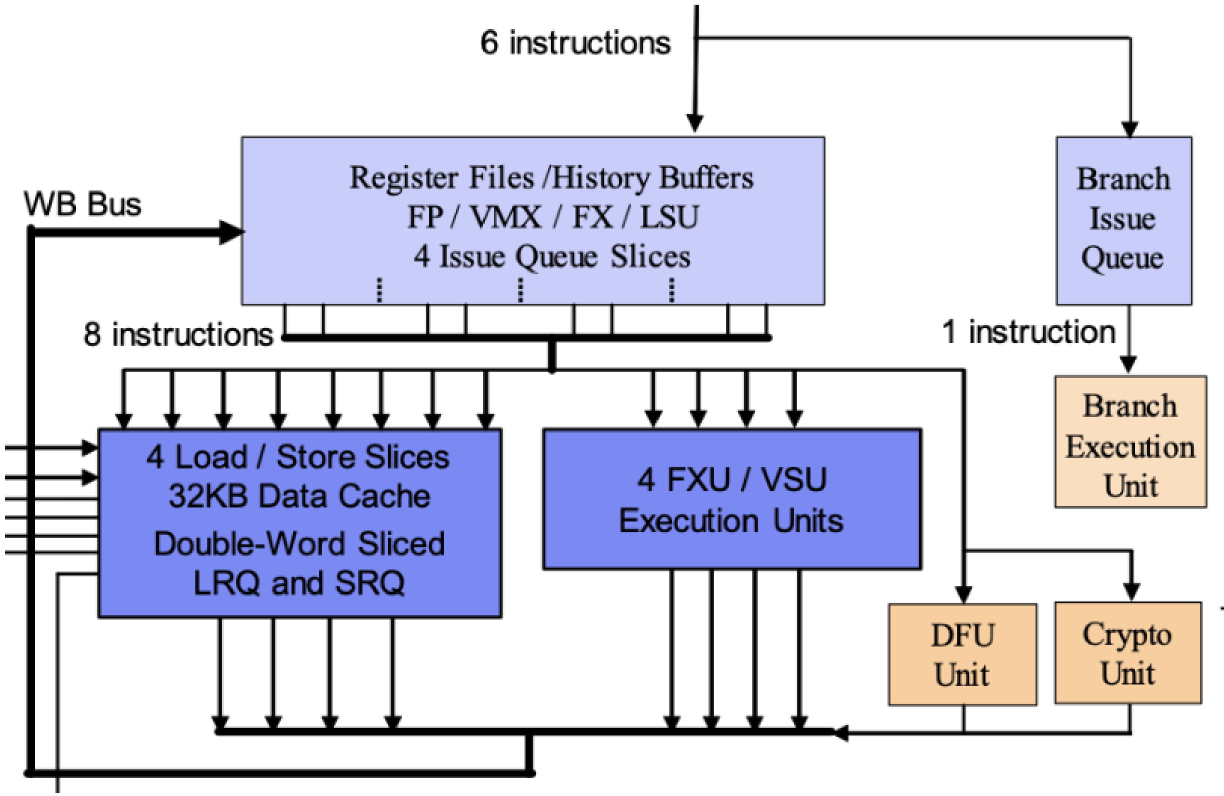
Core Frontend



- 8 instructions placed into IFB every cycle
  - IFB capacity: 96 instructions
- 6 instructions decoded concurrently every cycle
  - Some instructions cracked into 2 or 3 instructions
    - Costs 2 cycles in the pipeline
- Instructions speculatively dispatched in-order

Core Backend

- History buffer & Register Renaming used for O-o-O execution
- 11 pipelined FU:
  - 4 LSU
  - 4 FXU/VSU
  - 1 Branch Execution Unit
  - 1 DFU
  - 1 Crypto Unit



Branch Prediction

- 4 branch history/prediction tables (8K entries with 2-bits)
  - 1 supplementary TAGE predictor
- Selector decides to either use local or global predictor
- Latencies:
  - 3 cycles for regular predictors
  - 5 cycles for TAGE predictor

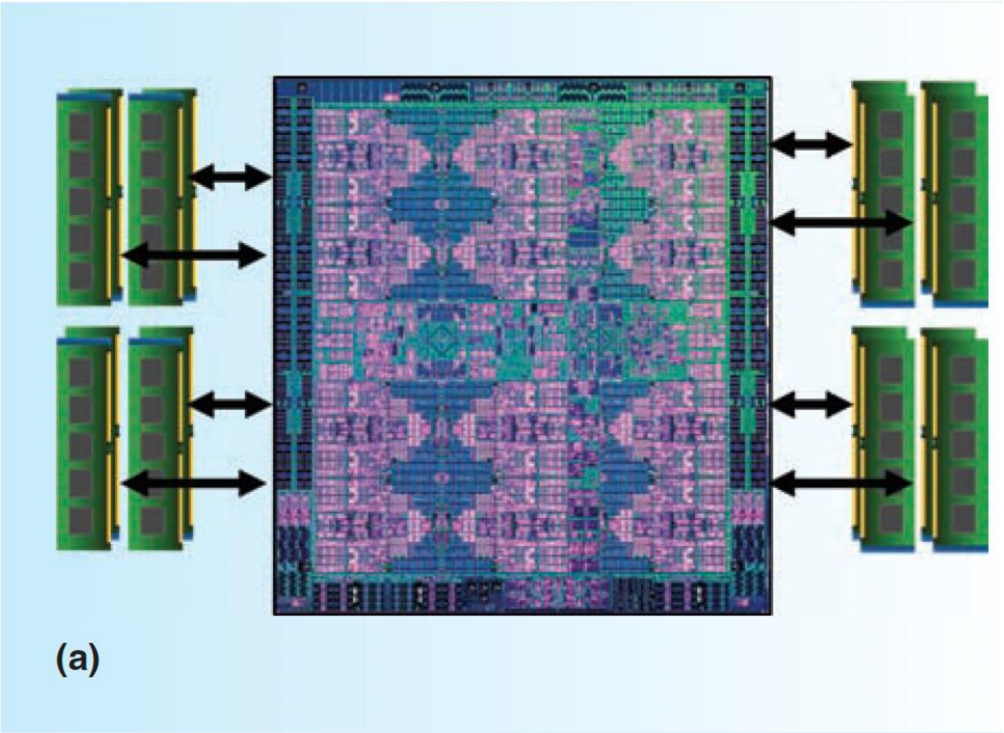
Load/Stores

- 4 reads, one write per cycle
  - If no conflict between the write and a read
- On an L1 D-cache hit, 4 cycle load-use penalty
  - 3 cycle penalty between load and a dependent op
- Out of order loads:
  - Total of 76 outstanding loads
  - Done through load reorder queues
    - Keeps track of out of order loads and watches for hazards
- Load-miss queue
  - Keeps track of loads that have missed the L1 D-cache
  - Merges multiple loads of the same cache line into one entry

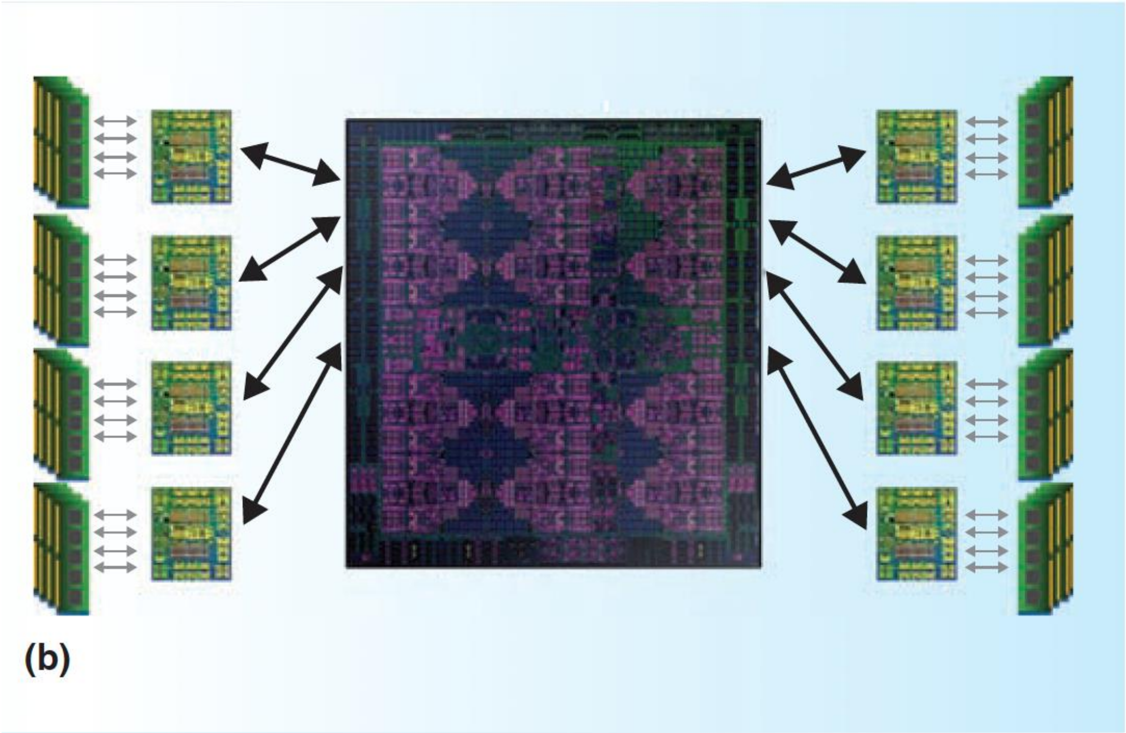
Data prefetch

- Adaptive prefetch mechanism
  - Adapt prefetch aggressiveness to increase performance based on prefetch consumption and memory bandwidth
- Prefetches and allocates ahead of demand into
  - L1 D-cache from the L3 cache.
  - L3 cache from memory.

Memory Hierarchy Overview



- Scale-out:
  - Direct attach
  - 8 DDR4 ports



- Scale-up:
  - Buffered memory attach
  - 8 buffered channels

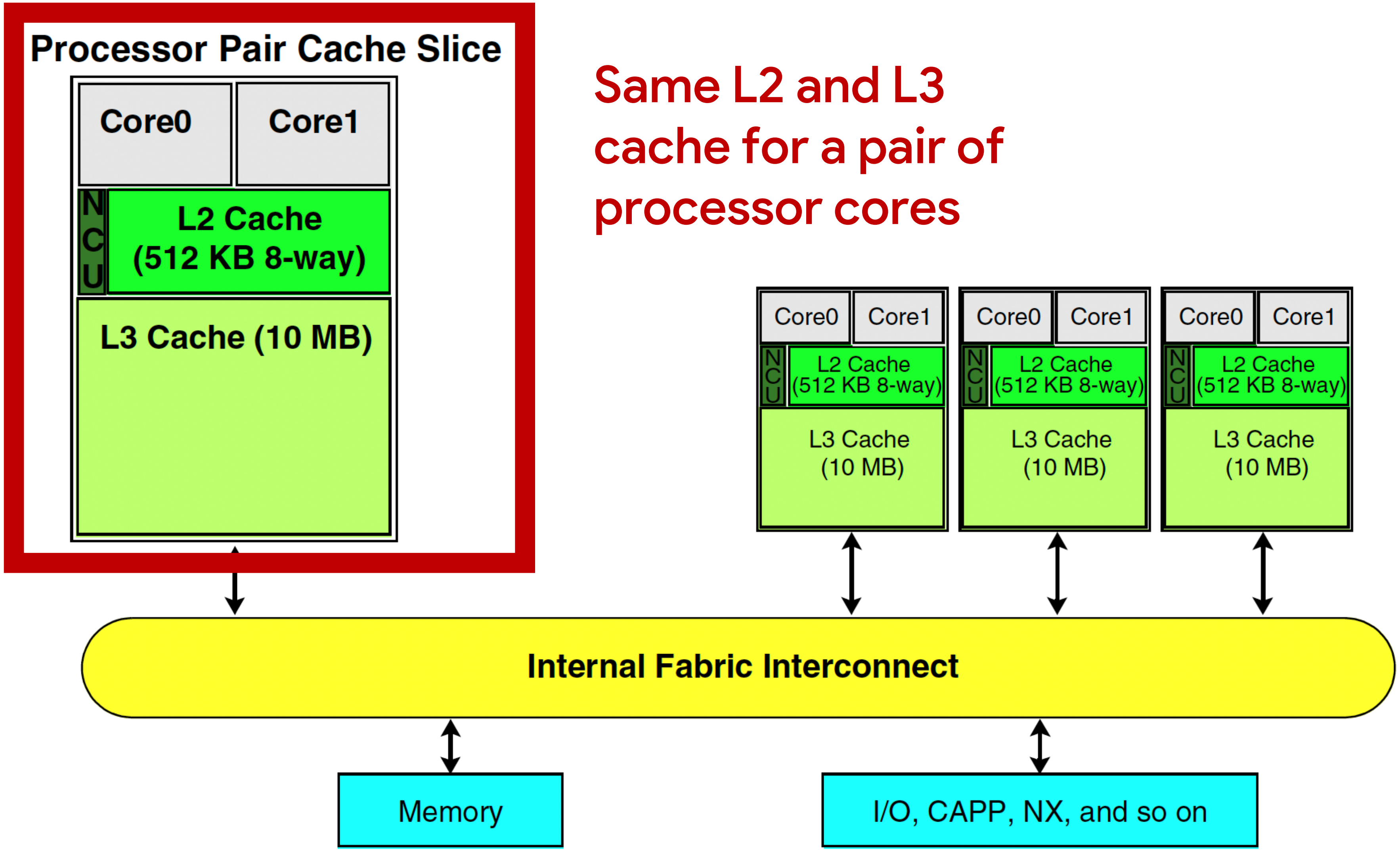
Cache Overview

- Three levels of cache hierarchy:
  - L1 (I-cache, D-cache),
  - L2,
  - L3
- Dynamically shared between different threads
- Cache line:
  - 128-bytes

L1 Cache

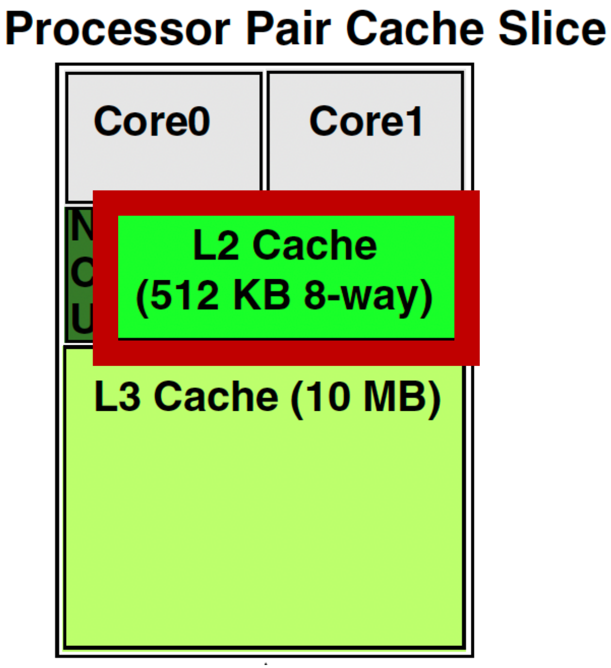
- Dedicated L1 cache for each core
- Separate I-cache and D-cache (each 32 KB)
- All cache lines in L1 are also present in L2
- 8-way set associative
  - Indexed with virtual address, but...
  - tag comparison on physical address -> no synonym hazard
- Dual banked
- Writethrough, no-allocate
- Pseudo-LRU replacement policy

L2 and L3 Caches (in SMT4 version)



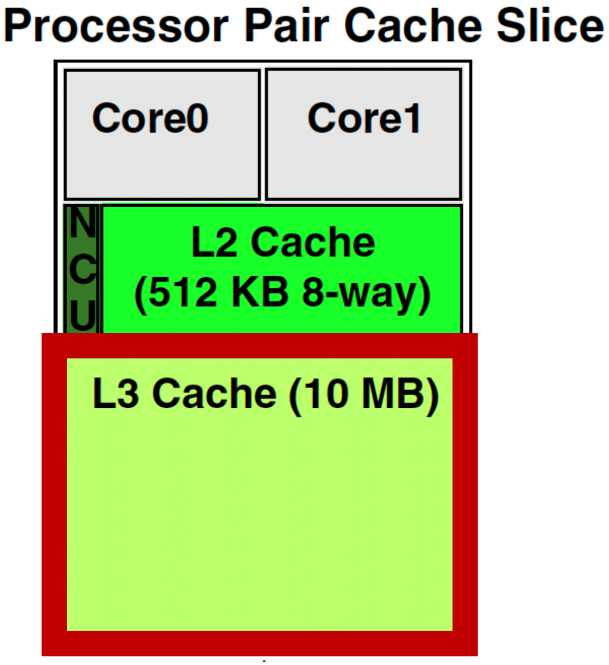
Same L2 and L3 cache for a pair of processor cores

L2 Cache



- 512 KB
- 128-byte line
- 8-way associative
- Write-back, write-allocate
- Maintains full-hardware coherence within the system

L3 Cache



- 10 MB
- 20-way associative
- 10 banks
- Victim cache
  - for L2 and L1 caches
- Maintains full-hardware coherence within the system

Multicore & Multithreading

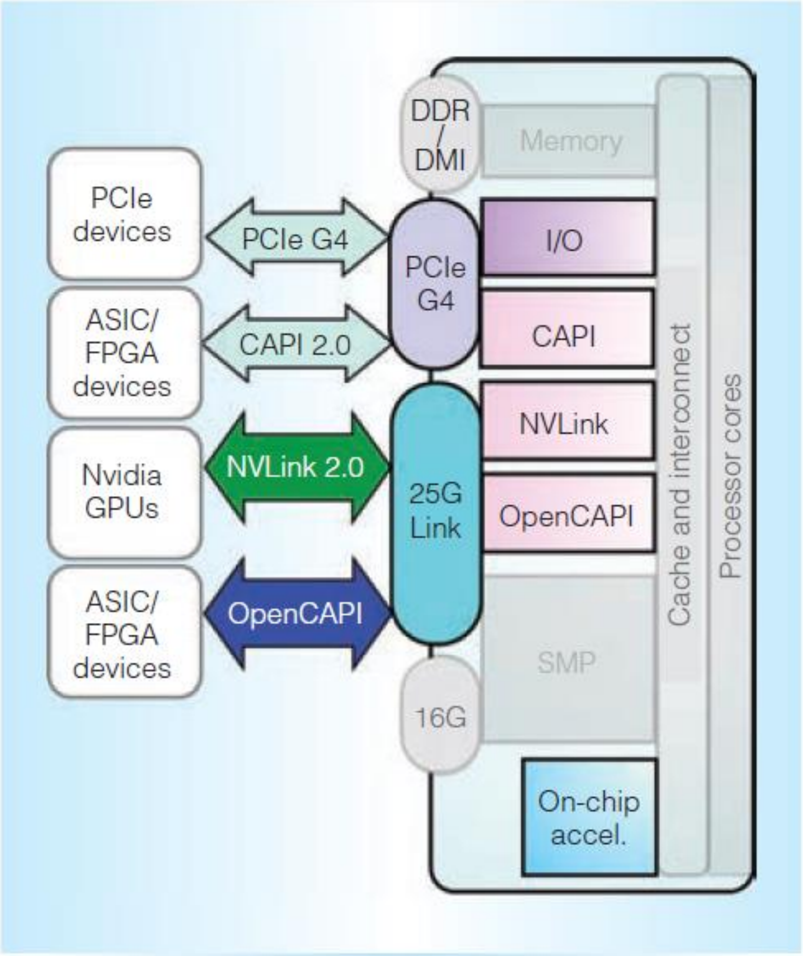
- Supports Simultaneous Multithreading
  - Up to 4 threads per core for SMT4 version
  - Shares core execution resource
- HW support for:
  - Thread prioritization
  - Balance work between threads
  - Forward progress

SMP Interconnect

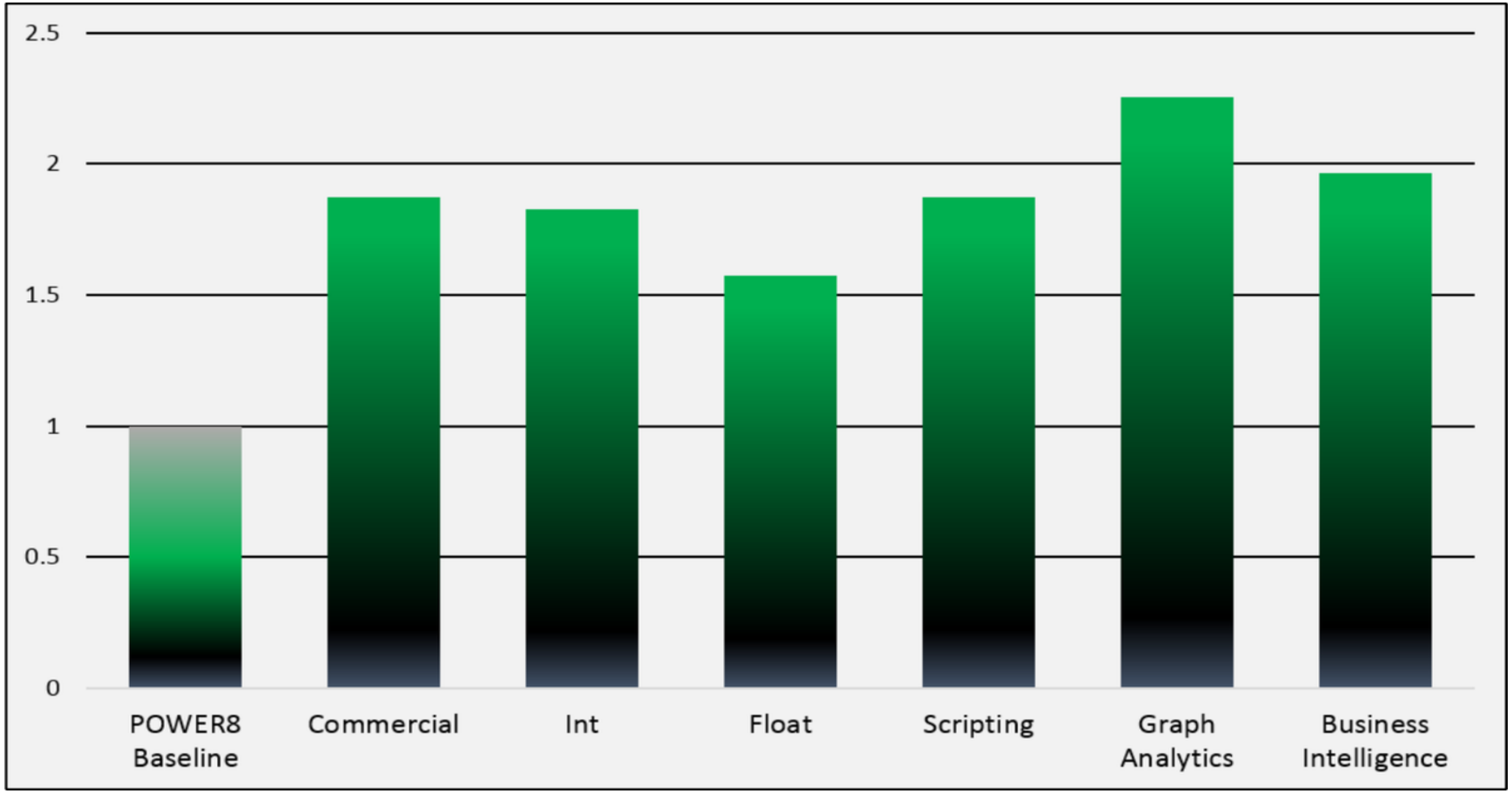
- Underlying hardware for cache-coherent multiprocessing
- Provides coherent and non-coherent memory access, I/O operations, interrupt communication, and system controller communication
- Integrated on the POWER9 chip (with 24 cores & on-chip memory system)
- Cache coherence is maintained by using a snooping protocol

Accelerators

- On chip accelerators:
  - GZIP & 842 compression/decompression engines
  - AES & SHA cryptographic hash engines
- Off chip accelerators:
  - CAPI & OpenCAPI interfaces:
    - Connect ASIC and field-programmable gate array (FPGA) devices
  - Nvidia NVLink 2.0 interface:
    - Connect Nvidia GPUs
  - Coherent memory sharing
    - Reduces overhead due to data interactions between CPU & accelerators



Performance



The graph represents a scale-out model of similar specs at a constant frequency.

References

- IBM Power 9 Processor User’s Manual
- IBM Power 9 SMT Deep Dive Summit Training Workshop – Brian Thompto
- IBM Power 9 Introduction Summit Training Workshop – Brian Thompto
- S.K. Sadasivam, B. W. Thompto, R. Kalla and W. J. Starke, “IBM Power 9 Processor Architecture” in IEEE Micro
- Power 9 Microarchitectures – IBM - WikiChip

Thank you!

Happy to take your questions

Backup slides

Address translation

- TLB:
  - 1024-entry, 4-way set-associative
  - Shared by the four threads
  - 4 KB, 64 KB, 2 MB, 16 MB, 1 GB, and 16 GB pages are supported in the TLB
  - Hardware-based reload (from the L2 cache interface; no L1 D-cache corruption)
  - Support for four concurrent table walks
  - Hit-under-miss is allowed
  - Binary LRU replacement policy