IBM Power9

Disha Agarwal, Minhyuk Park, Akash Kothari

About IBM Power9

- Power9 chip architecture is built for technical/HPC workloads, hyperscale, analytics, cognitive, and predictive machine learning applications.
- The Power9 architecture is open for licensing and modification by the OpenPower Foundation members. Sierra supercomputer built by Lawrence Livemore National Laboratory is a good example.
- Supported by FreeBSD, IBM AIX, IBM i, Linux, RHEL, Debian GNU/Linux and CentOS.

Pipelining

- Reduced pipelined design.

Register to register operations	11 stages
Load store operations	13 stages
Floating point operations	17 stages

- Longer instructions are broken down into smaller instructions.
- In-order dispatch of up to 6 internal operations (2 branches, 6 non branch) into 5 distributed issue queues per cycle.
- Out-of-order issue of up to nine operations (4 load/store, 4 add/sub/mul/div/mv , 1 branch).
- Register renaming.

Execution Units

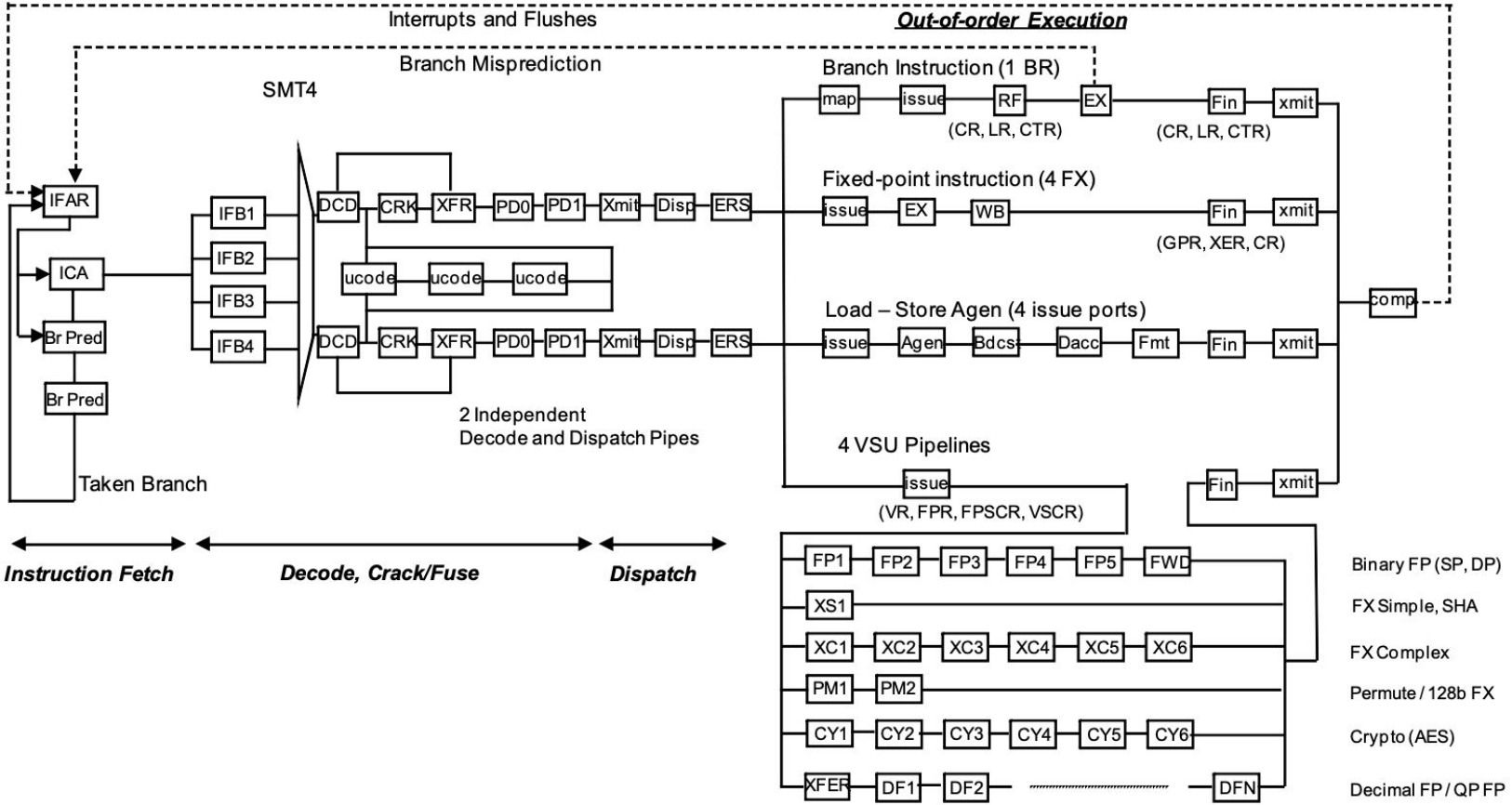
Load/Store Units	4
Symmetric 64-bit VMX execution units	4
DFU	1
Crypto	1
Branch execution unit (BR)	1

Pipeline Structure

- Master pipeline
  - speculative in-order instructions to the mapping, sequencing, and dispatch functions.
  - ensures an orderly completion of the real execution path (Mispredicted paths are thrown away).

- Execution unit pipelines
  - Out-of-order issuing of both speculative and non-speculative operations.
  - Independent from the master pipeline.

Pipeline Structure



[2]

Branch Prediction

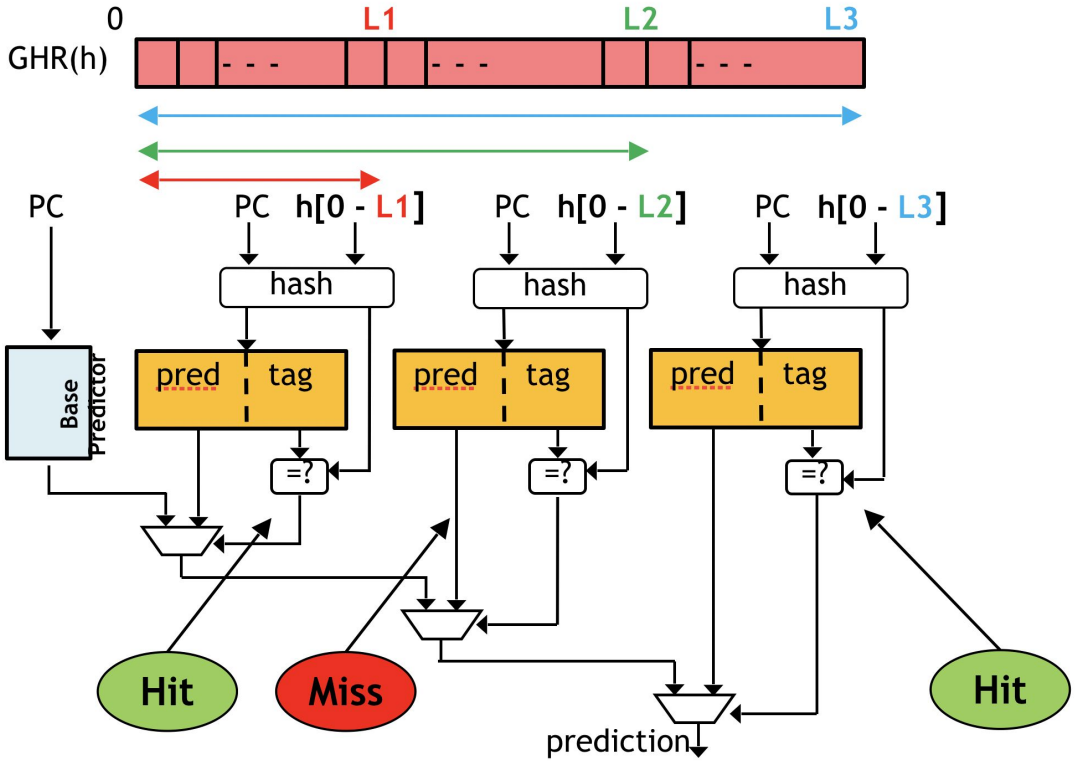
- Predict up to 8 branches per cycle.
- Prediction support for branch direction and branch target addresses.
- Number of support for predicted taken branches: 40 (single threaded), 20 (2 threaded) and 10 (4 threaded).
- 4 branch predictors: local predictor, global predictor, selector and local selector (8K entries x 2-bit).
- Branch predictors backed up by TAGE predictor.

TAGE Branch Predictor

- Tagged Geometric History Length (TAGE).
- Uses multiple tagged tables and different global history lengths.
- Set of history lengths forms a geometric series.

[4]

TAGE Branch Predictor



[4]

Memory Hierarchy

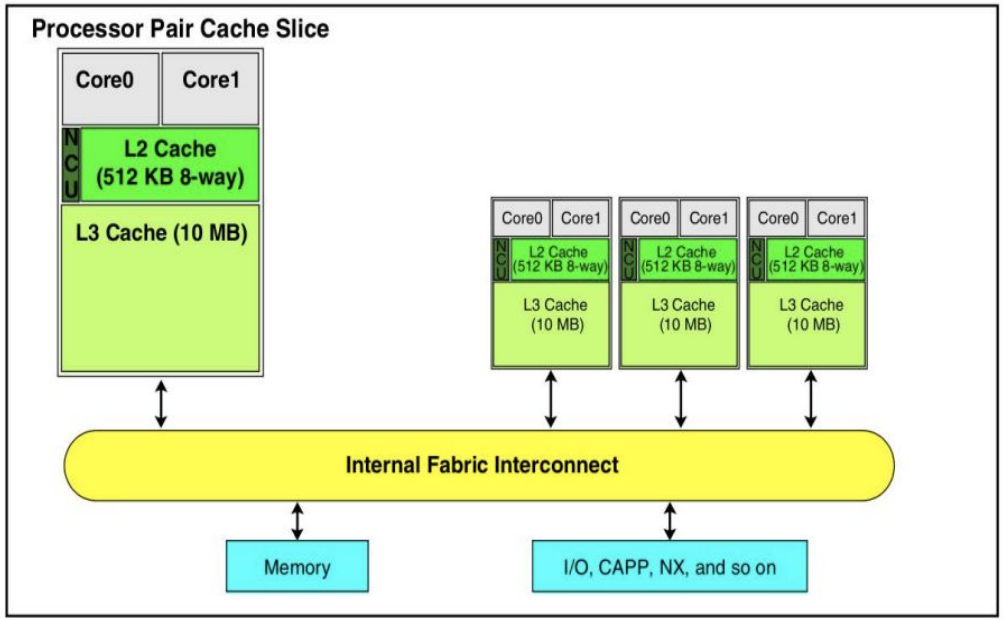
- L1 Cache
  - 32 KiB, 8-way set associative
  - 128-byte cache line
  - Separate Data and Instruction Cache
  - Pseudo-LRU replacement policy

- L2 Cache
  - 512 KiB 8-way set associative
  - 128-byte line
  - Dual Banked
  - LRU replacement policy

- L3 Cache
  - 120 MiB eDRAM
  - 10 MiB/core pair 20-way associative
  - Per-entry state-based replacement policy

Basic Multicore Overview

- 24 cores per chip.
- L1 cache is not shared.
- L2 cache is shared between cores pair.
- L3 cache is victim cache for L2.



[1]

Data Prefetching

- 8 independent data streams capable of striding up or down
  - N-stride detection.
- Prefetches and allocates ahead of demand into the L1 D-cache from the L3 cache.
- Prefetches and allocates ahead of demand into the L3 cache from memory.
- Software-initiated prefetching.
- Adaptive Prefetching:
  - Predicts the confidence based on stream and program history.
  - Assigns confidence levels to each prefetch request.
  - Allows the memory controller to drop less confident prefetches.
  - Identifies phases of program execution where prefetching might be more effective.
  - Receives feedback from the memory controller.

Address Translation

- 1024-entry, 4-way set-associative TLB per cluster.
- 4 KB, 64 KB, 2 MB, 16 MB, 1 GB, and 16 GB pages are supported in the TLB.
- The TLB entries are shared by the 4 threads as long as the entry belongs to the logical partition running on the core.
- Hit-under-miss is allowed in the TLB.
- Support for four concurrent table walks.
- 2 outstanding table-walks per cluster and TLB hits-under-miss is allowed.
- Both software and hardware TLB management is allowed.
- LRU replacement policy.

Scale out vs Scale up

- Scale out variant is for single or dual-socket setups.
- Scale up variant is for servers with quad + socket setups.
- Supports a large memory capacity and throughput.

Simultaneous Multithreading

Three supported modes:

- ST (single thread)
- SMT2 (2 concurrent threads)
- SMT4 (4 concurrent threads)

SMP Interconnect

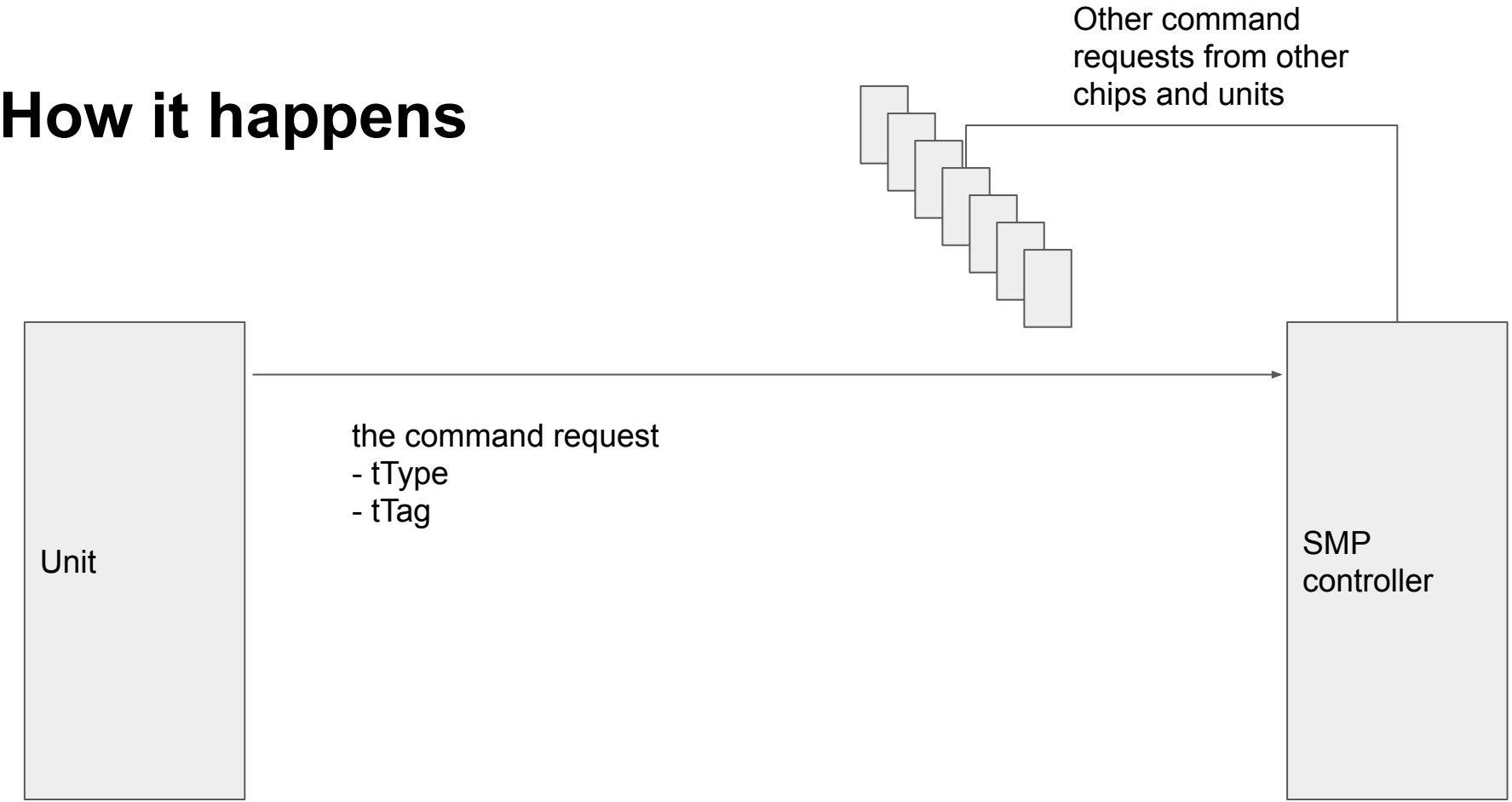
SMP (Symmetric multiprocessing) Controller is in charge of:

- Coherent and non-coherent memory access
- I/O operations
- Interrupt communication
- System controller communication

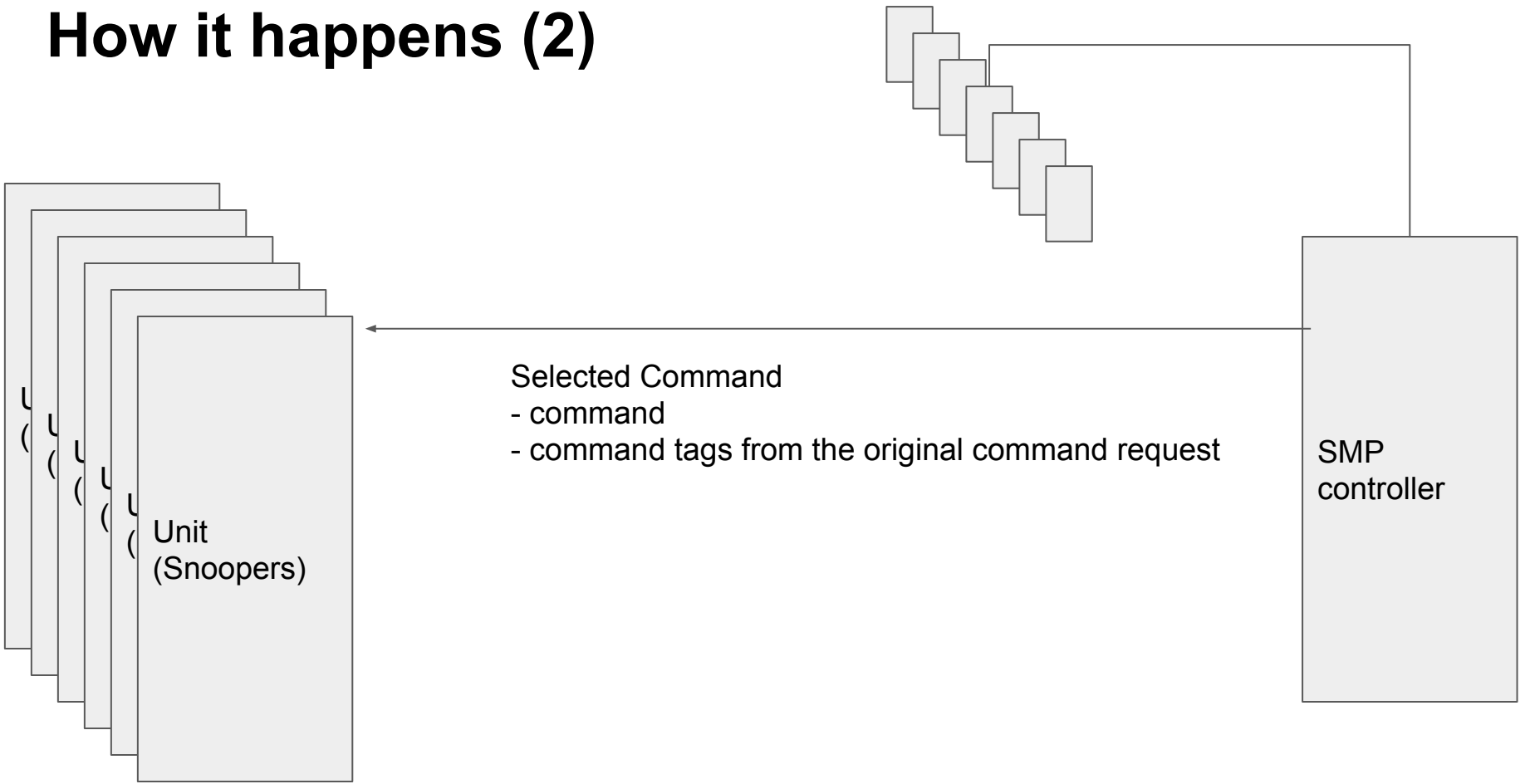
Cache Coherence

- Snooping protocol.
- Uses multiple levels of snoop filters.
- If cache coherence is impossible, then the command must be re-issued at a higher scope (local node, remote node, group, and vectored group).

How it happens

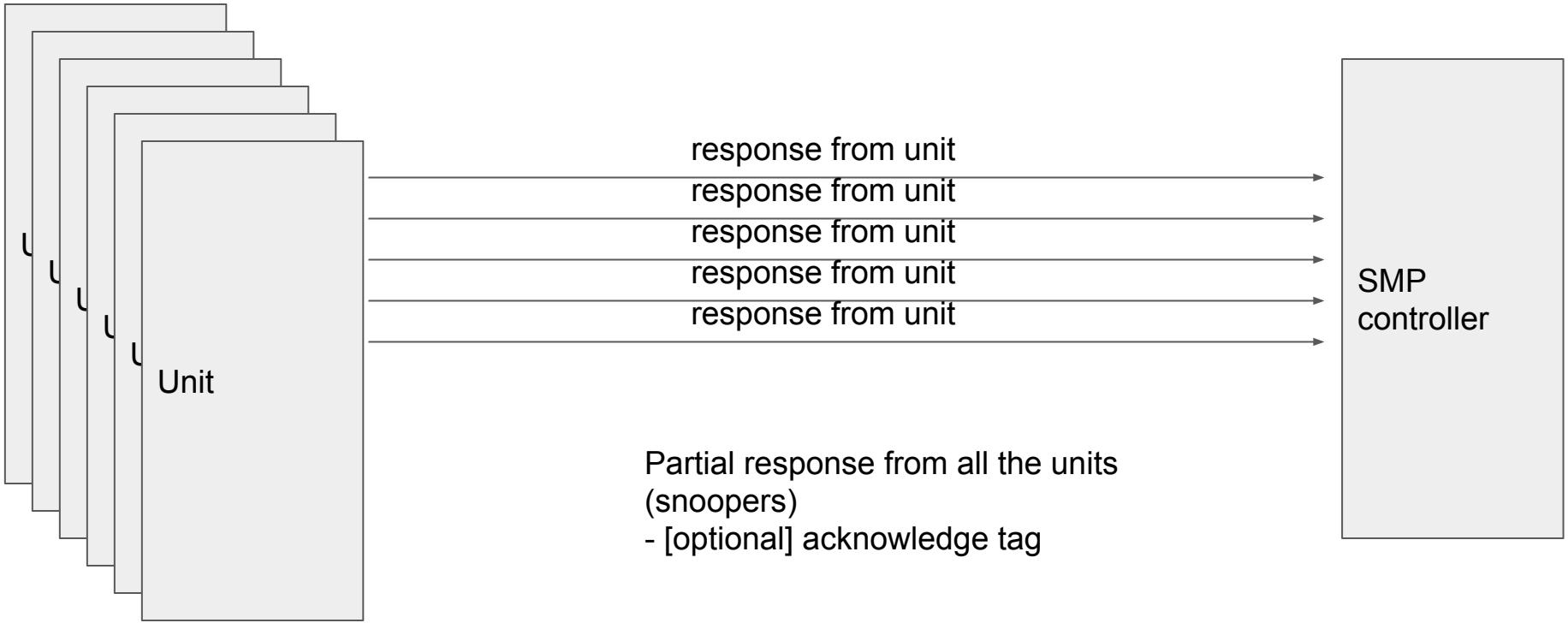


How it happens (2)



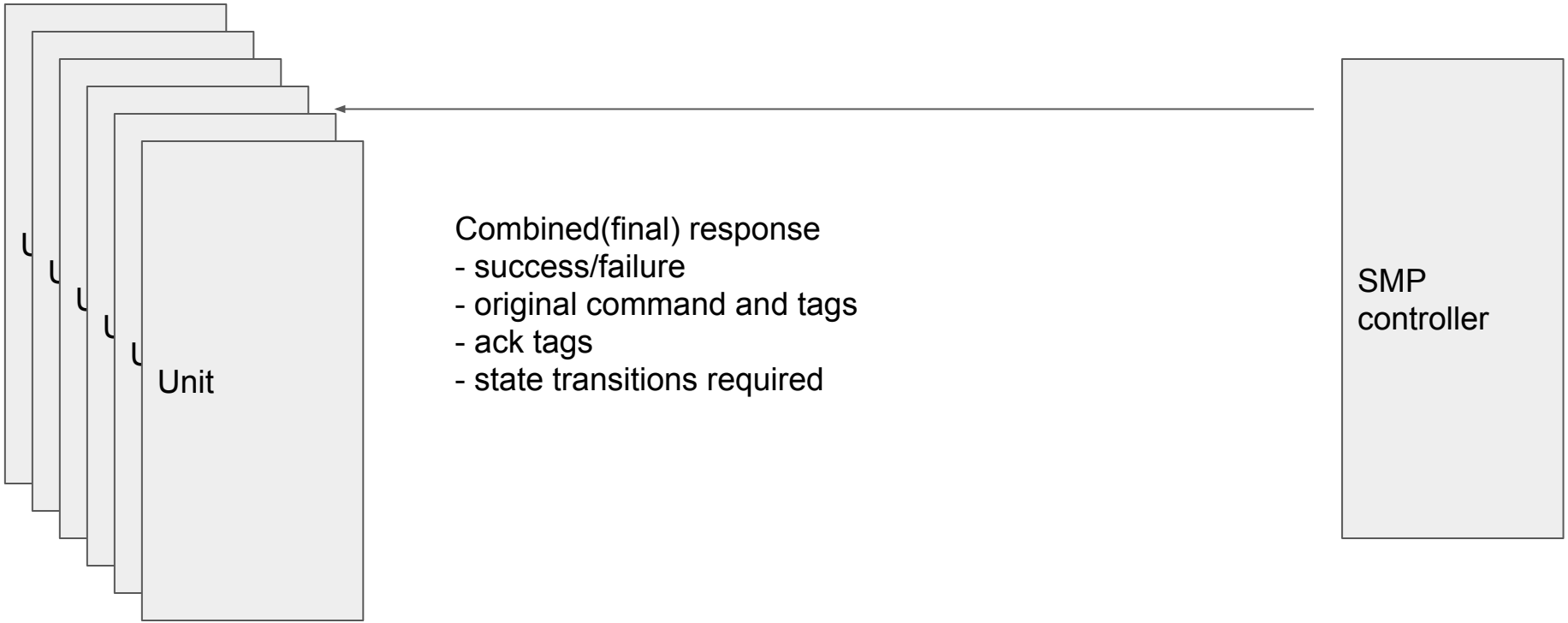
How it happens (3)

(tSnoop time interval later)



Partial response from all the units
(snoopers)
- [optional] acknowledge tag

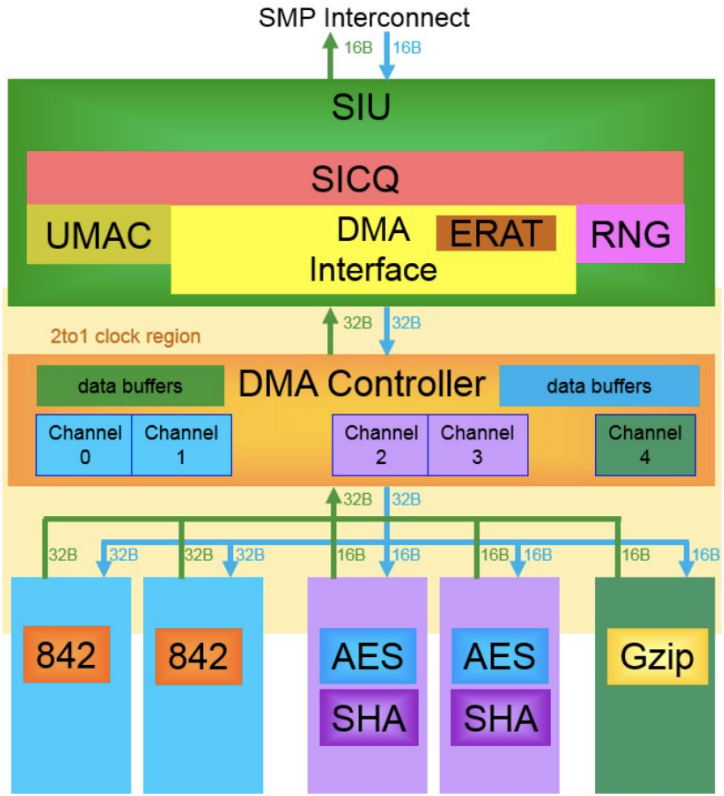
How it happens (4)



Nest Acceleration (NX)

Cryptographic Engines

- 2 AES engines
  - 128, 192, 256 bits keys
- 2 SHA engines
  - SHA-1, SHA-256, SHA-512 modes.
  - HMAC supported for SHA.
- Random Number Generator (RNG)
- Two 842 compression/decompression engines
- 1 Gzip compression/decompression engines



[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. A. Seznec, P. Michaud, “A case for (partially) tagged Geometric History Length Branch Prediction”, Journal of Instruction Level Parallelism , Feb. 2006.
4. TAGE Slides: https://ece752.ece.wisc.edu/lect09-adv-branch-prediction.ppt

Backup slides