NVRAMOS 2011 Fall
Operating System Support for
Next Generation Large Scale NVRAM

Accelerating Database Machine using In-storage Processing inside SSDs

Sungchan Kim
Chonbuk National University, Korea

In collaboration with
Hyunok Oh, Hanyang University, Korea
Chanik Park, Samsung Electronics, Korea
Sangyeun Cho, University of Pittsburgh, U.S.A
Sang-won Lee, Sungkyunkwan University, Korea

Trends in Large-scale Data Processing

- Data-centric computing
  - Cloud computing, map-reduce, scientific data, analytics, search engine
  - Data stored in company, public internet, and home is doubling every month
    - Several TBs/sec data to be processed
  - Key operations
    - sequential scan / filtering / sorting / grouping / hashing …
- What implications on computing paradigm?

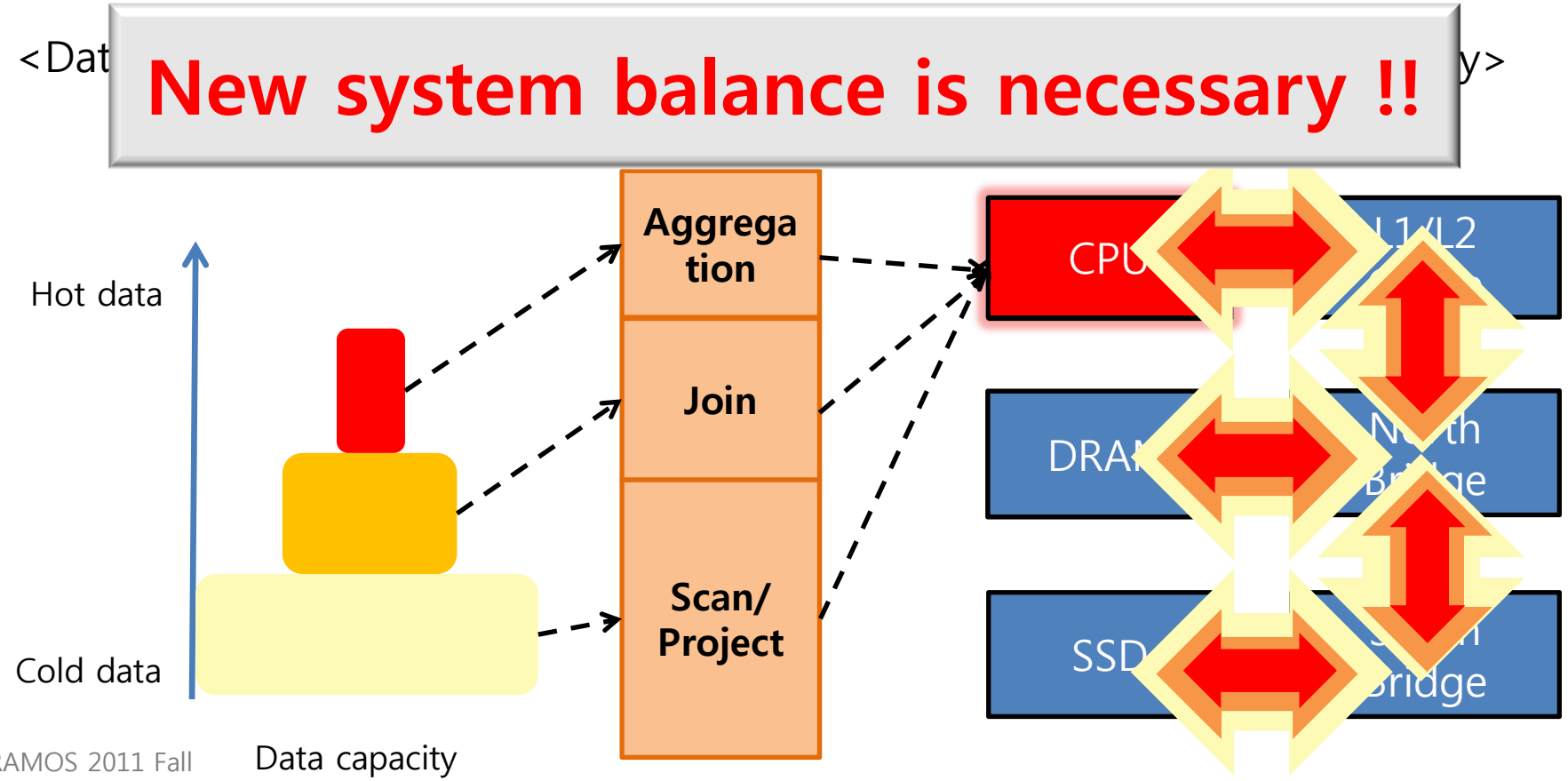
NO MORE Conventional Computing

- No data locality
  - Conventional memory hierarchy may not work
  - E.g. SCAN operation in DB
- No complex processing logics
  - Complex host CPU-based processing may be inefficient
- Then, new computing paradigm?
  - Flash SSDs are computers
  - In-Storage Processing inside flash Solid-State Disk(SSD)s

What is *In-Storage Processing* (ISP)*?*

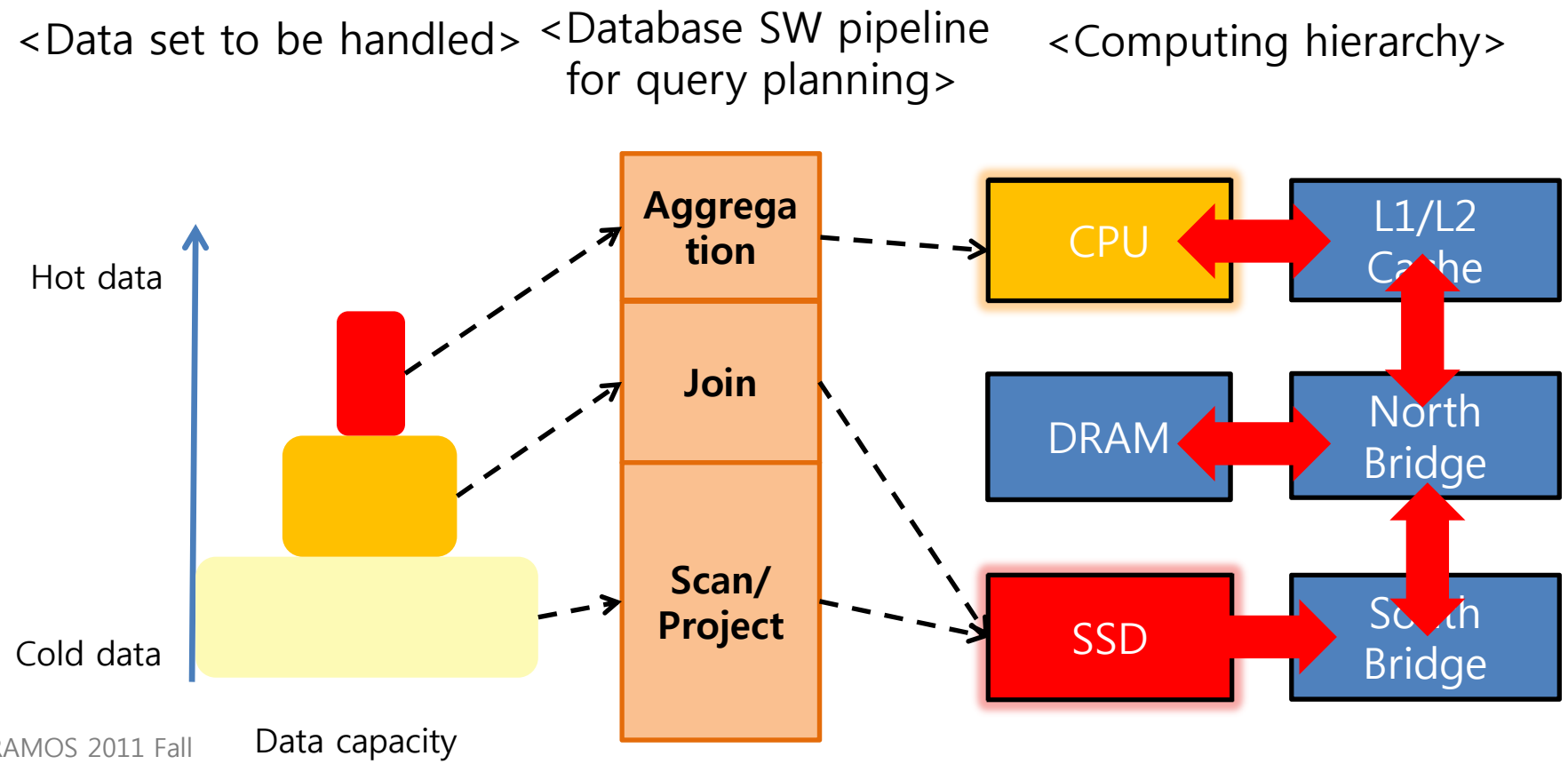
Conventional Host Processing

- Current limitations
  - Bandwidth wall in conventional multiprocessor system in handling data intensive applications
  - Datacenter PUE (Power Utilization Efficiency)

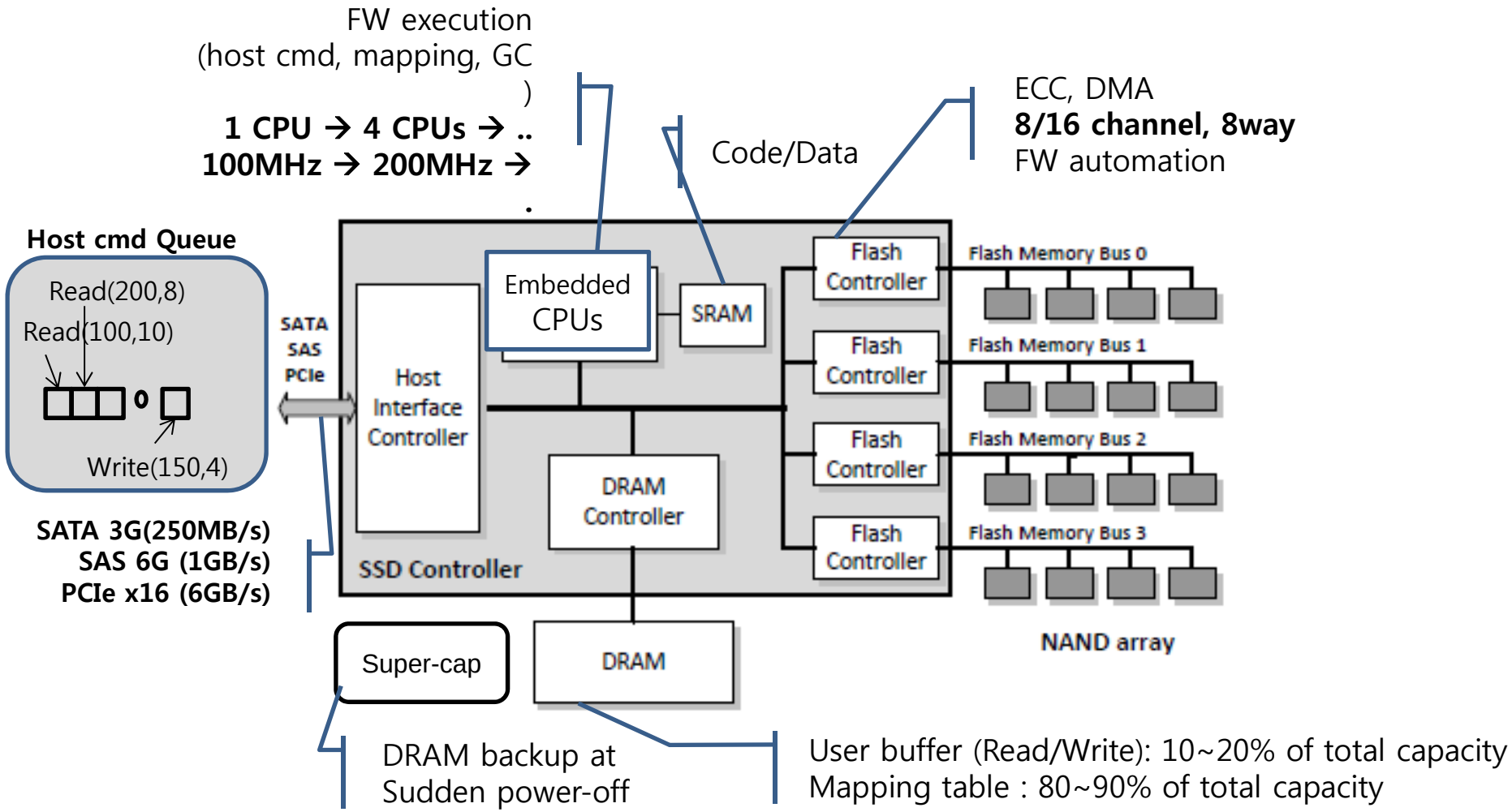


In-Storage Processing (ISP)

- Offloading host processor by moving (a part of) computations to storage medium
- Significant reduction of amount and latency of data transfer
- Unlimited I/O bandwidth

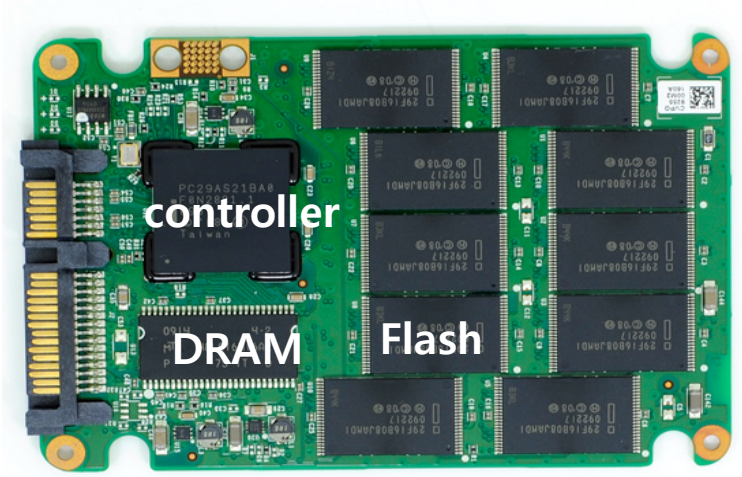


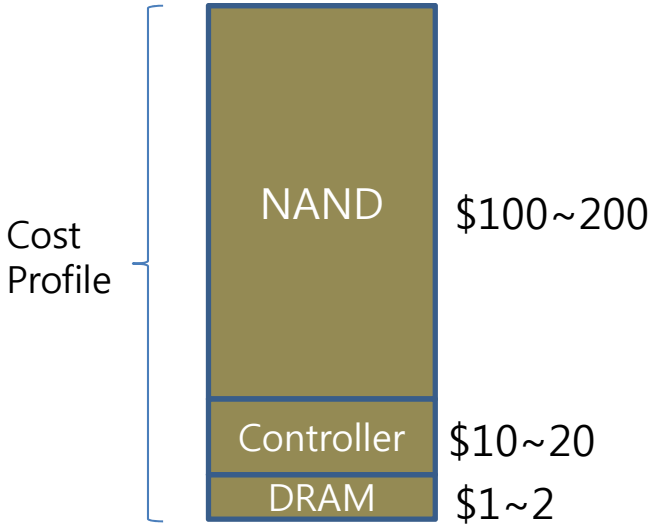
Typical SSD Architecture



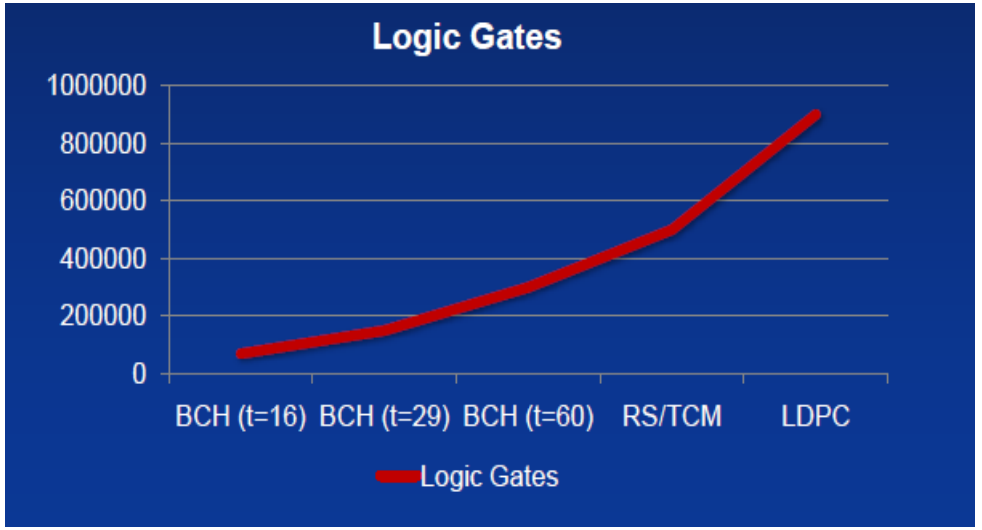
Enabling Technologies: SoC Technology

- Integration of massive computing elements
- NAND: a dominant cost factor
  - 80~90% depending on the density
- ECC: the dominant area factor
  - Affected by NAND technology.
  - CPU and simple logic (e.g., compare) seems to be small adders.

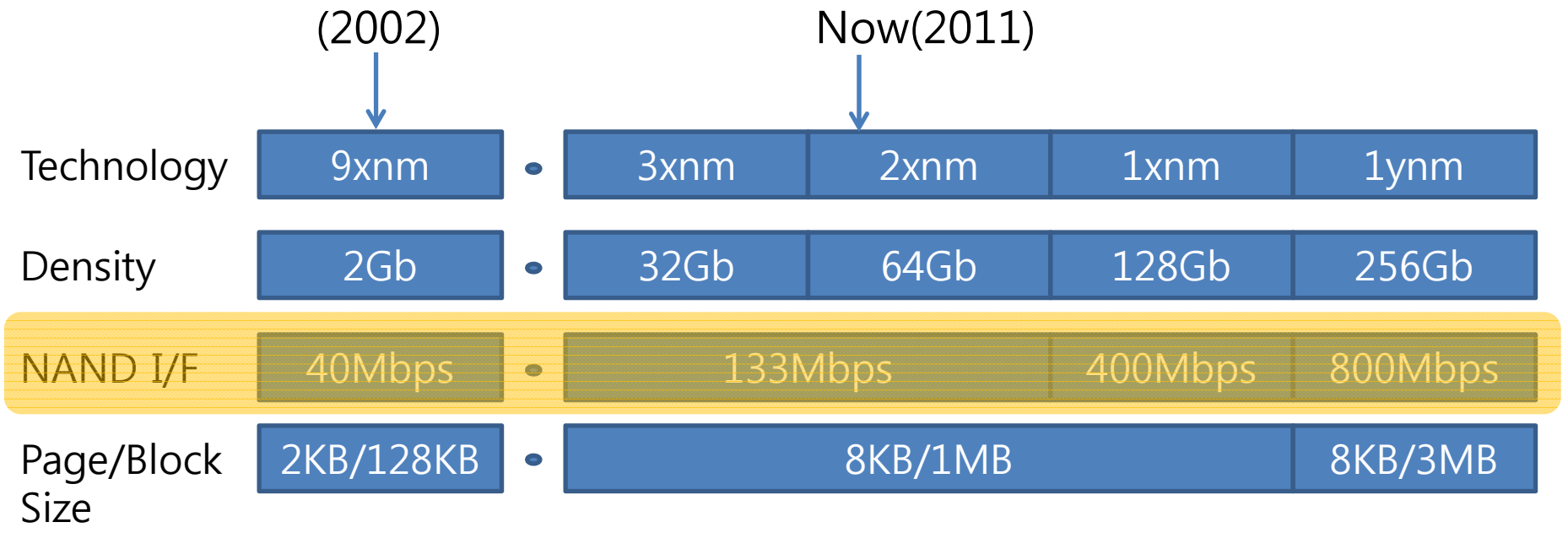


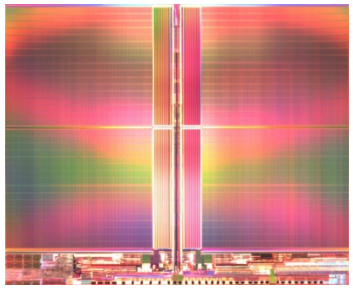


ECC HW Cost

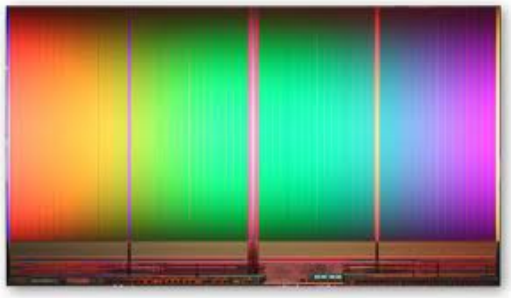


Enabling Technologies: High Speed NAND





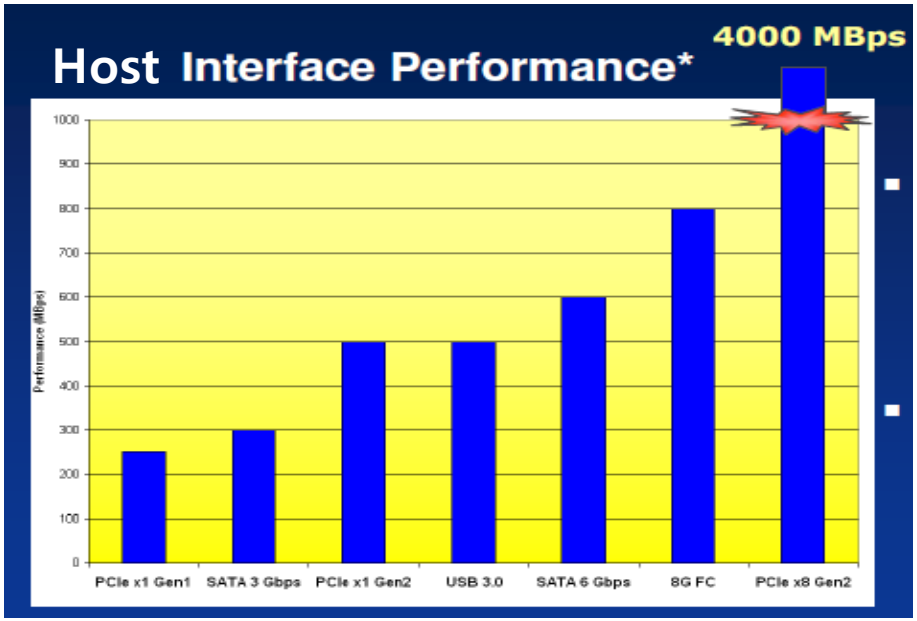
2 plane



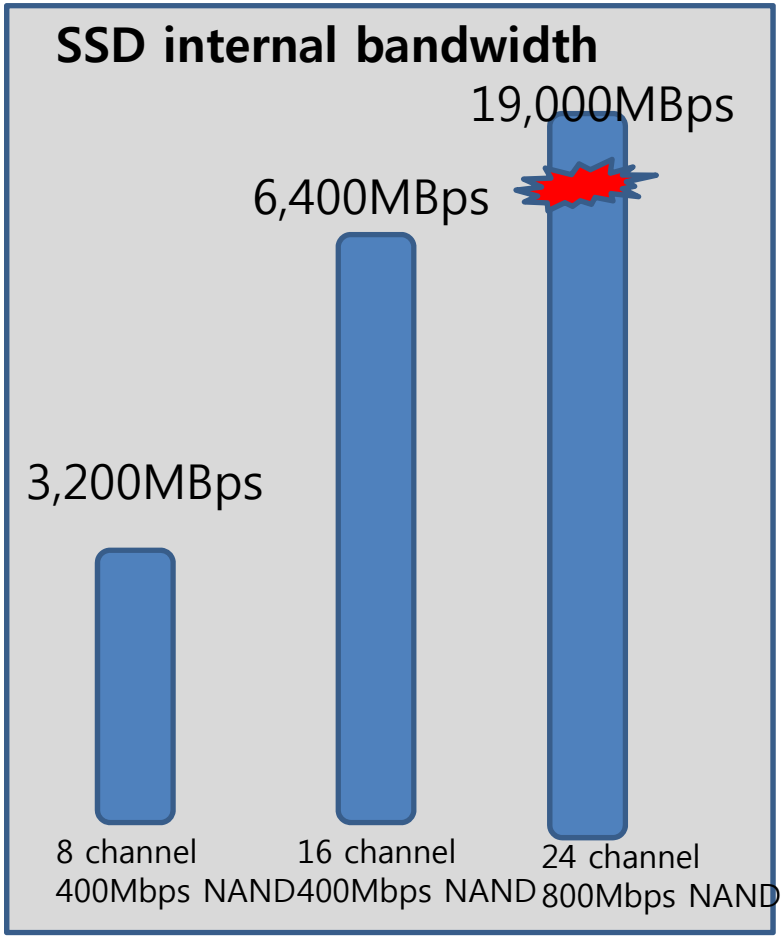
4 plane

Enabling Technologies: Increasing SSD Internal Bandwidth

- The internal bandwidth of SSD can surpass that of host interface.
- Translating internal bandwidth to data processing rate
  - Up to 19 Giga operations per second of compare operations



(IDT Flashmemorysummit'10)



Architecture for *In-Storage Processing* (ISP)

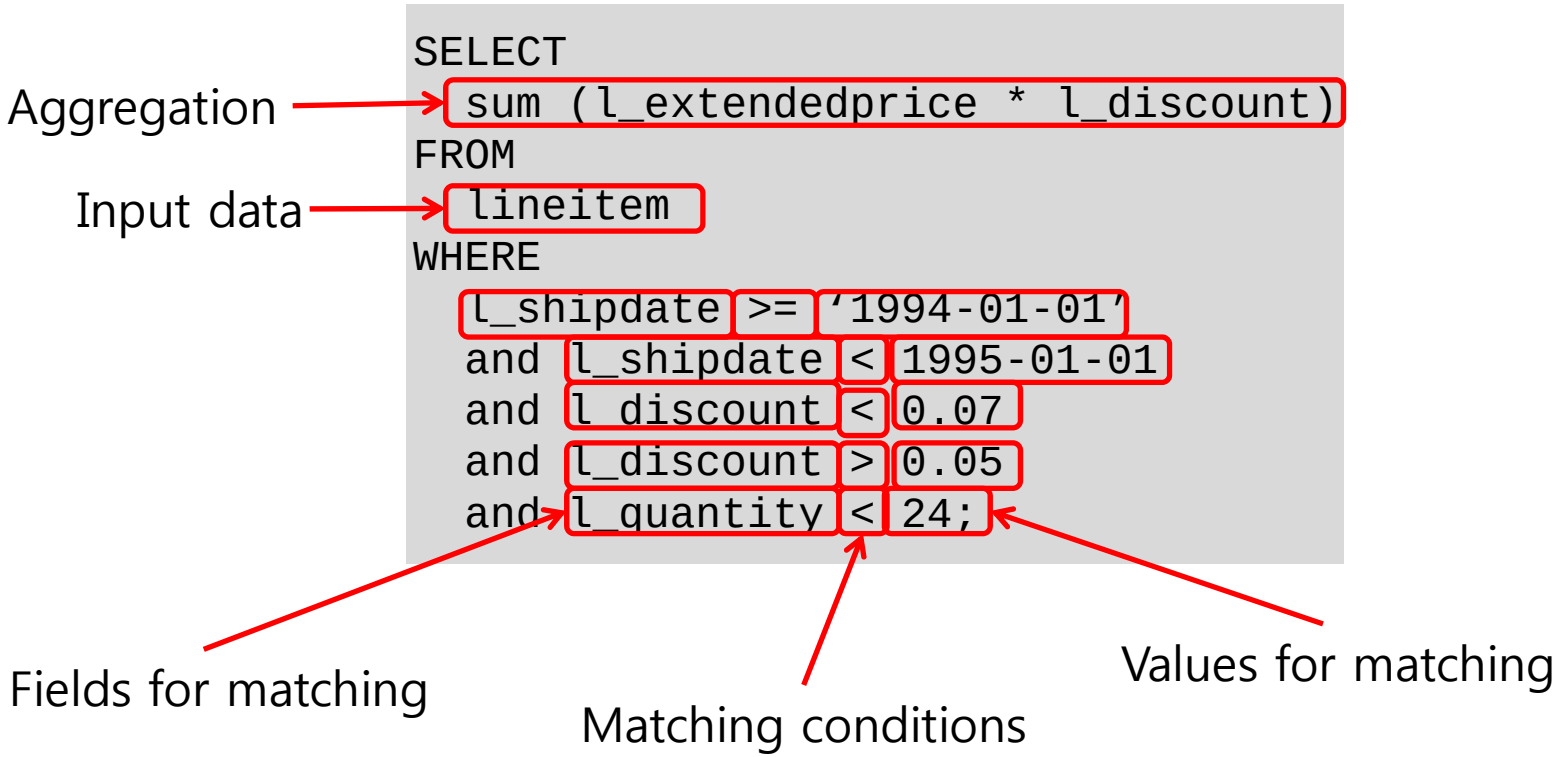
Why SCAN as Target Operation?

- Low data locality
  - Only a small portion of record is used
- Parallelizable
  - Multiple records can be scanned simultaneously
- Simple operation
  - Hardware realization is feasible

- Reduction
  - Aggregation / Low scan selectivity
    - Below 1% in our experiments based on TPC-H query

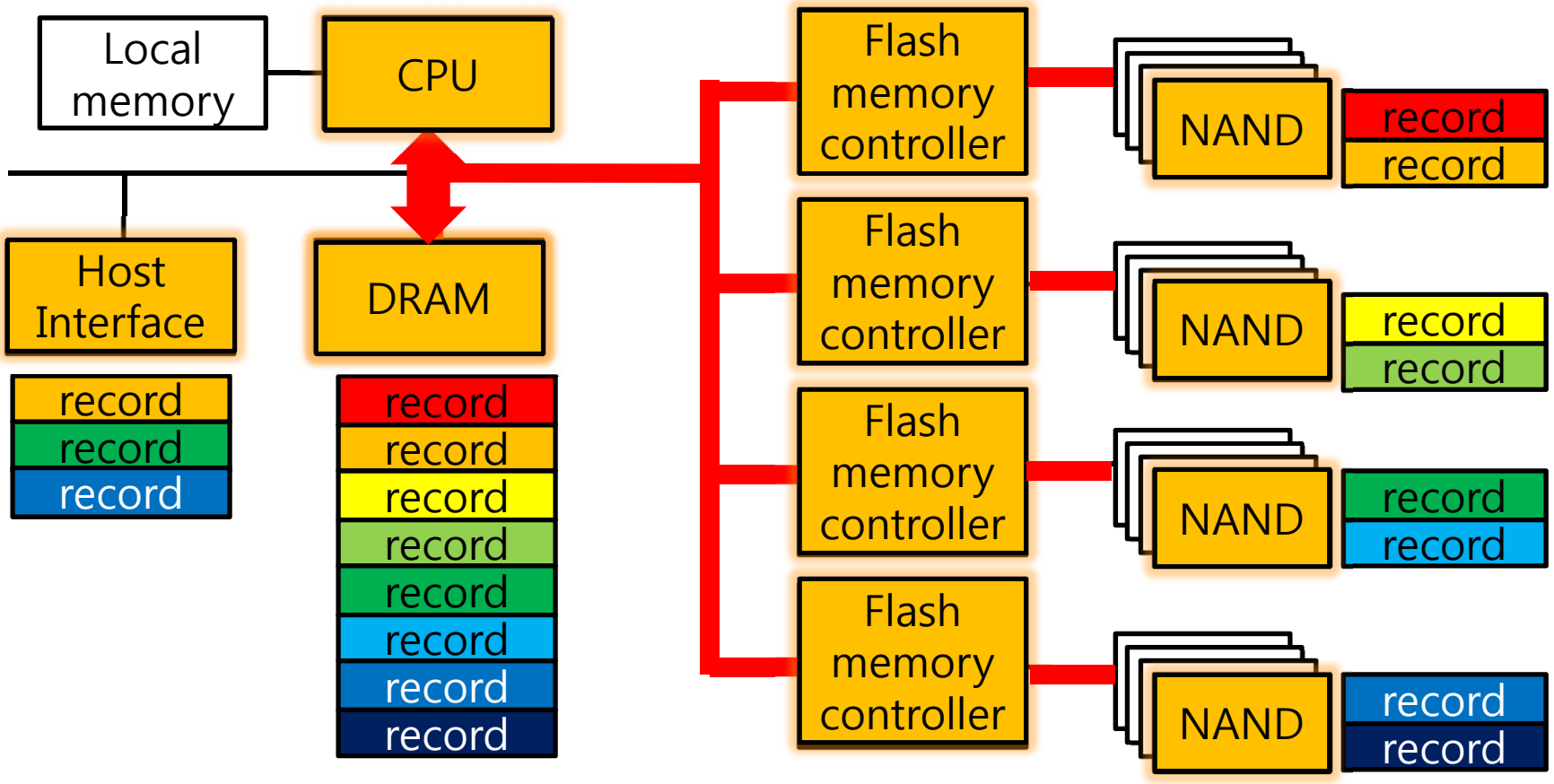
Example of SCAN Query

- Simplified Q6 in TPC-H benchmark



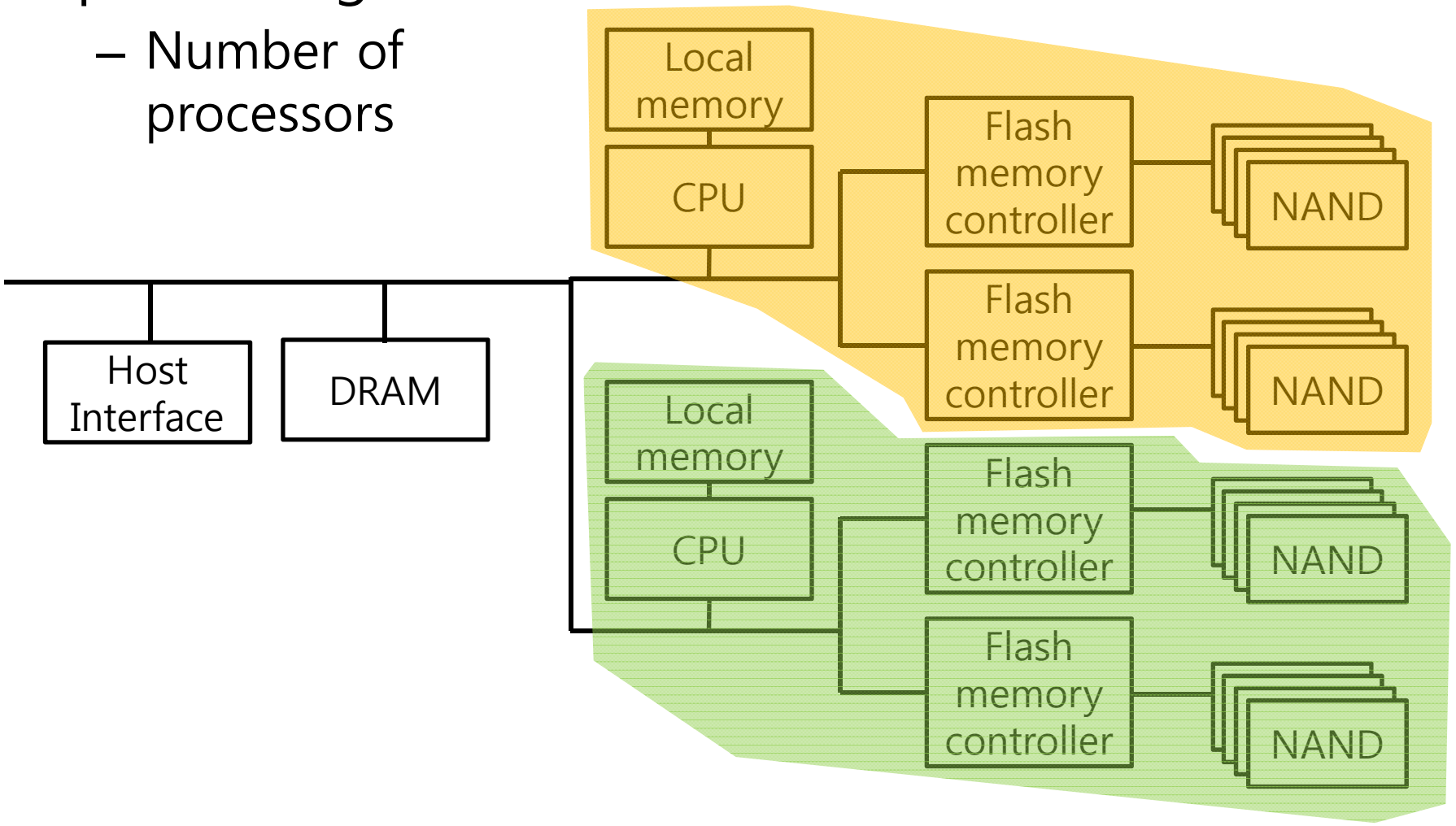
Baseline ISP Architecture

- Computation by an embedded CPU
  - Main performance bottleneck



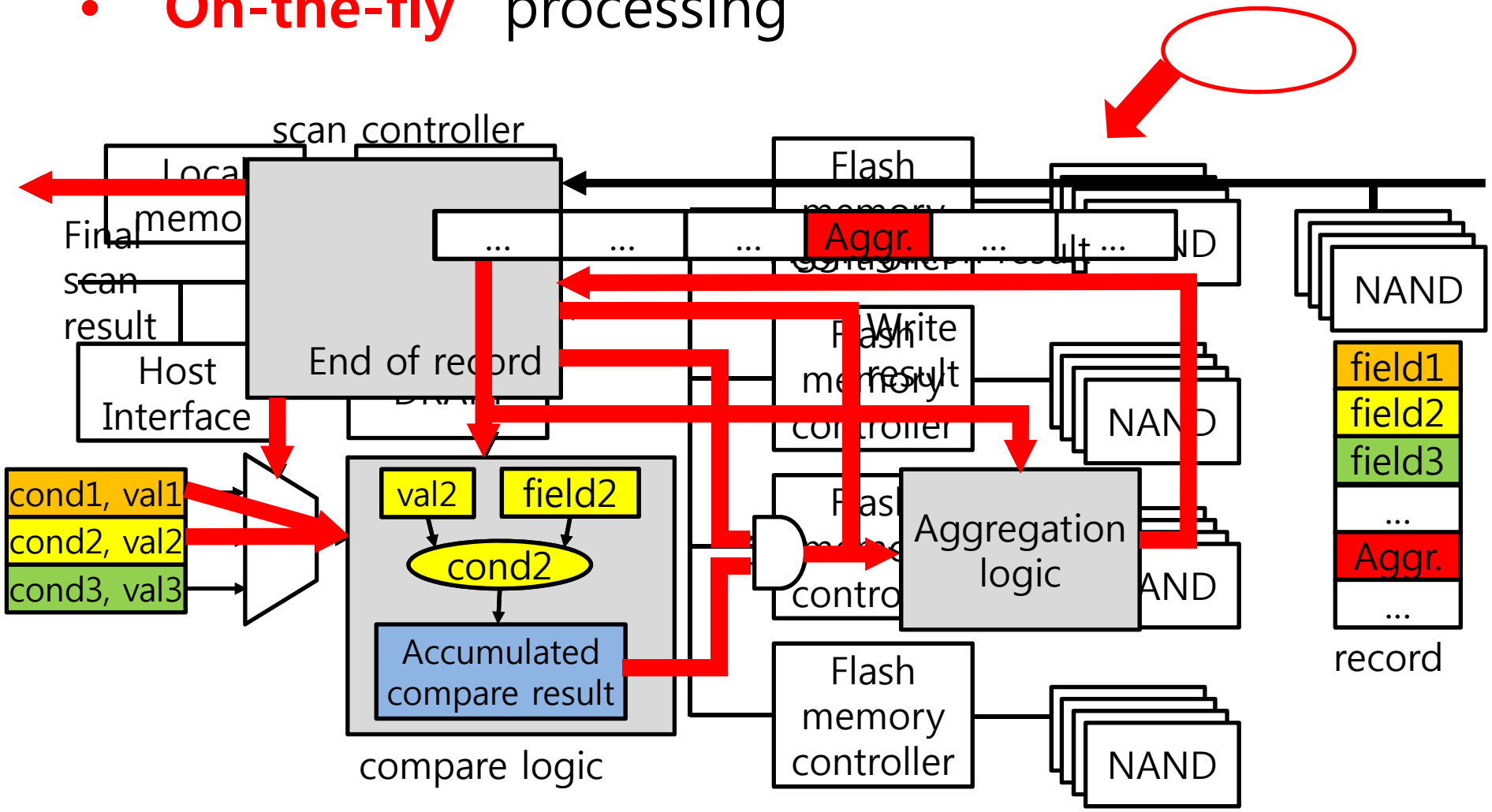
Multiprocessor-based ISP Architecture

- Dedicated processors for FMC-wise parallel processing
  - Number of processors



HW-accelerated ISP Architecture

- Parallel scans at each of FMCs
- **"On-the-fly"** processing



Evaluation

Analytic Model for Evaluation

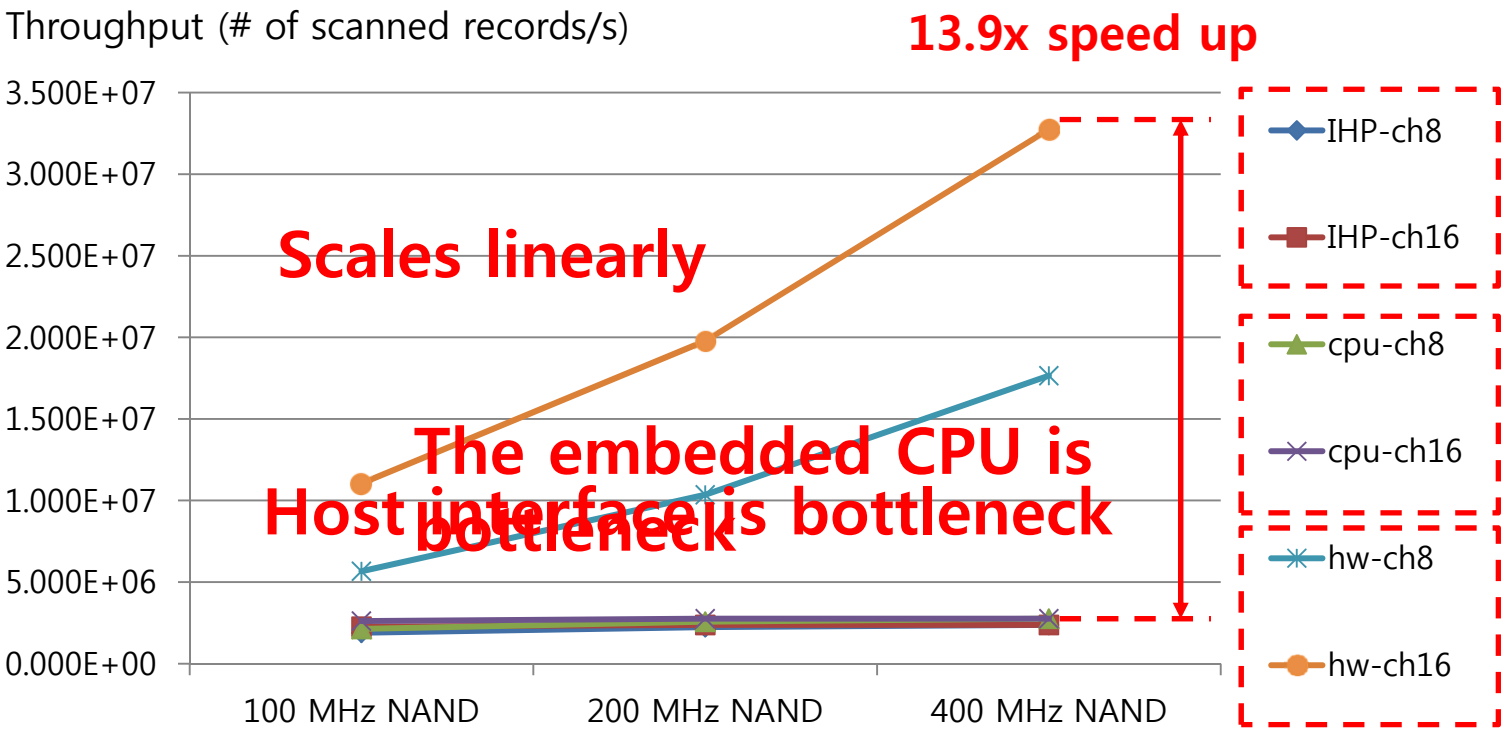
- Estimation of scan execution time
  - In-Host Processing / Baseline / HW-accelerated ISP
- Modeling accuracy
  - Comparison with cycle-accurate simulation model
  - Used query: Q6 in TPC-H benchmark

```
SELECT
  sum (l_extendedprice * l_discount)
FROM
  lineitem
WHERE
  l_shipdate >= '1994-01-01'
  and l_shipdate < 1995-01-01
  and l_discount < 0.07
  and l_discount > 0.05
  and l_quantity < 24;
```

	Baseline	HW-ISP
Model (cycles)	297282	16827
Simulation (cycles)	317446	16984
Error (%)	6.4 %	0.9 %

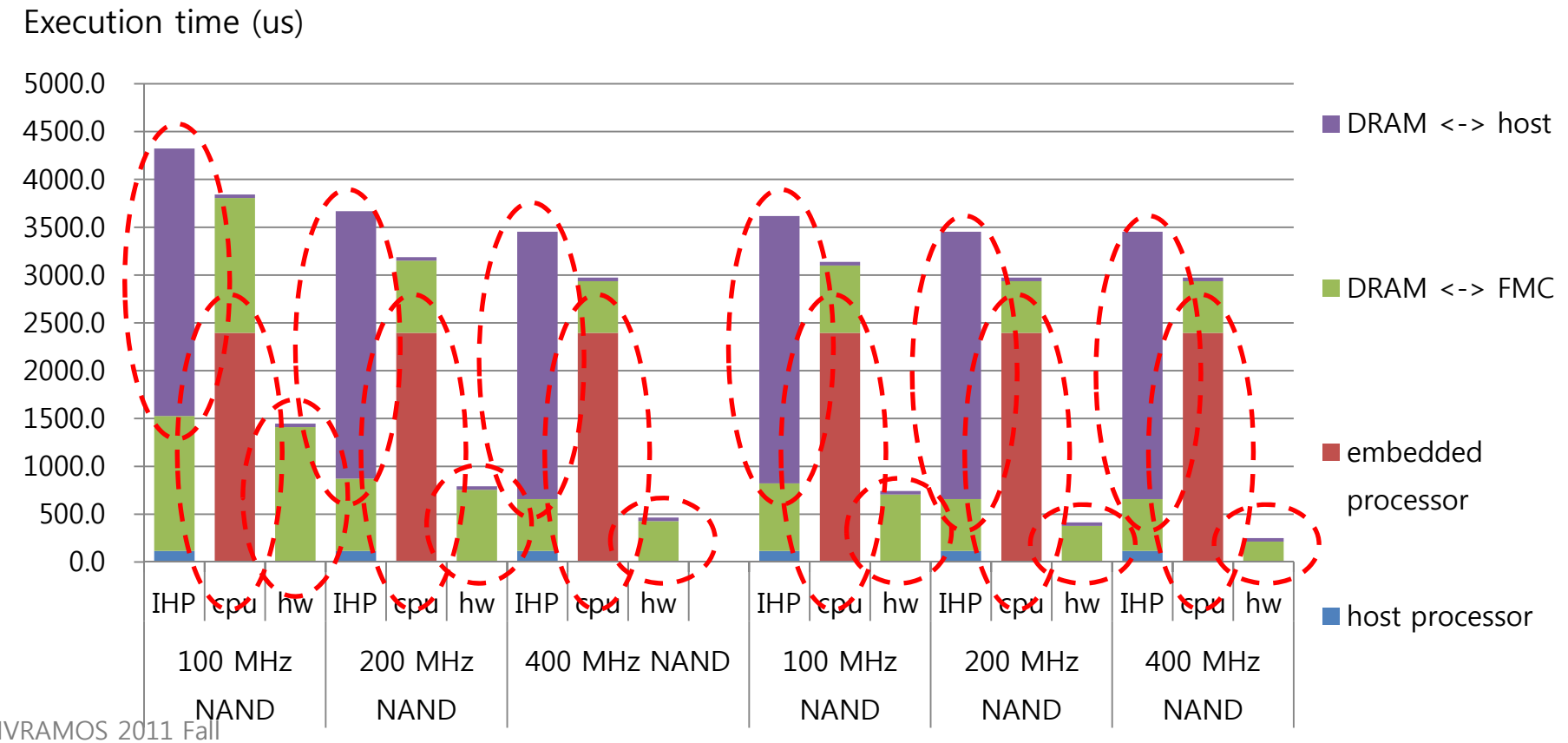
Throughput Evaluation

- Comparison of In-Host Processing and two ISP methods varying
  - Number of NAND channels: 8 or 16 channels
  - NAND interface speed: 100, 200, 400 Mbps
- Fixed host interface: SATA 2.0 (3Gbps)
- Low scan selectivity of 1 %



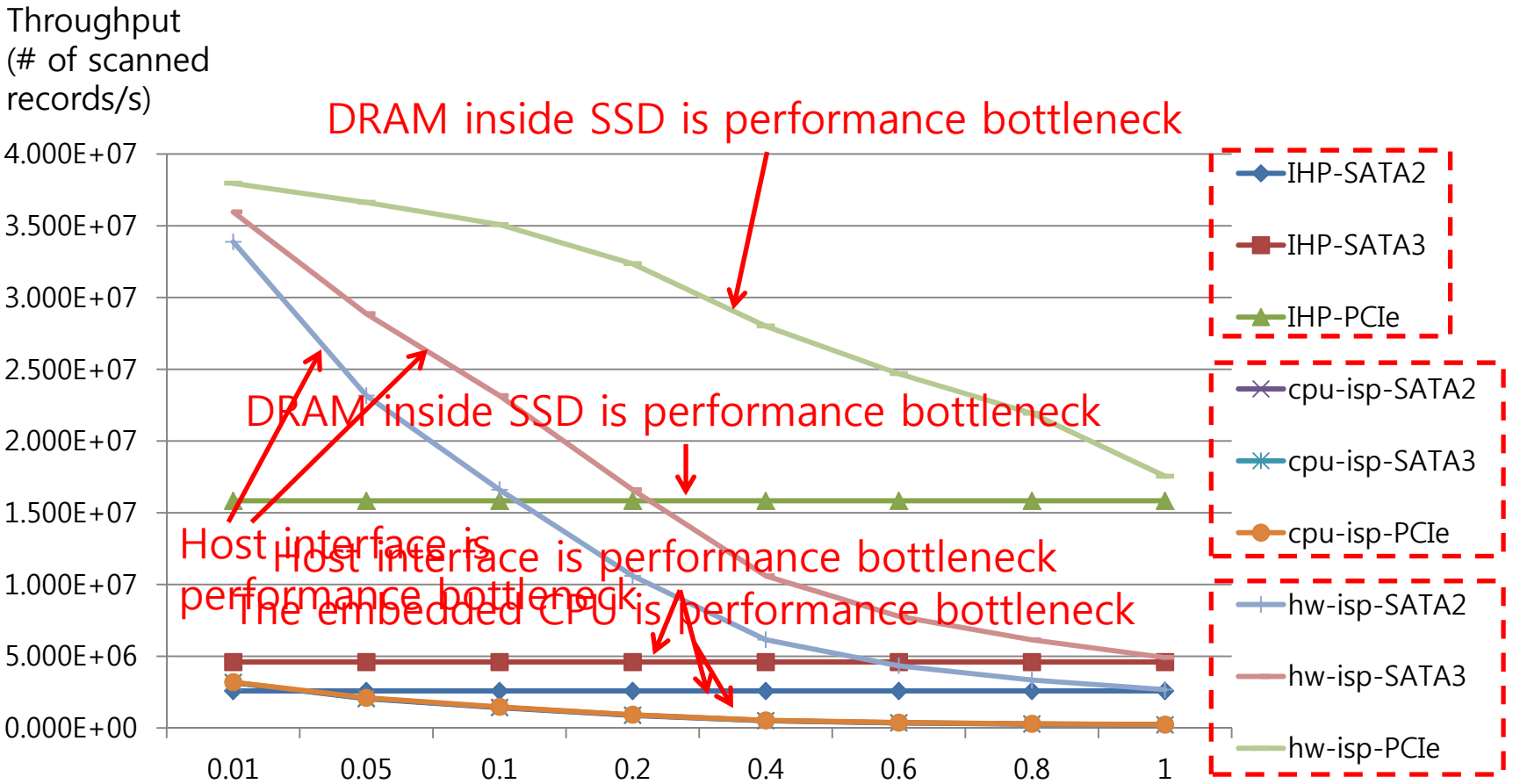
Where is Performance Bottleneck?

- In-Host Processing: data transfer
- Baseline ISP: embedded processor
- HW-ISP: **NAND-bounded performance**



Impact of Selectivity and Host Interface

- Host interface
  - SATA2 (3 Gbps), SATA3 (6 Gbps), PCI-e (64 Gbps)



Energy Consumption Evaluation

- Another key benefit of ISP
- Emulation of HW-ISP on a real SSD platform
  - Comparison based on actual measurement

Processing method	Normalized Energy consumption
ISP (modified firmware)	**0.142**
IHP (conventional)	1.000

Previous Efforts for ISP

- Database machine (1970s~1980s)
  - Accelerated operation with special purpose hardware per head, track, or disk
- Active disks (1990s)
  - Disk array with low cost embedded processors
  - Tries to offload host CPU's workload with the excessive computing power of the processors on disks
- Limitations
  - Limited bandwidth of disk media itself
  - Faster and faster commodity CPUs
  - No driving force in market
    - New storage interface, changes in software stacks
    - c.f. Oracle + Sun

Summary

- In-Storage Processing as next generation data-centric computing paradigm
  - DO NOT bring data to computation
  - BRING computation as close as to data
- We showed that
  - Significant performance/energy benefit potential
  - Difference performance bottleneck points compared to the previous ISP approaches
- Future work
  - More DB operations (join, sorting…)
  - Evaluation on real SSD platforms

For more details, see

→ Fast, Energy Efficient Scan inside Flash Memory SSDs, International Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architecture 2011.

Thank You !!

Q&A