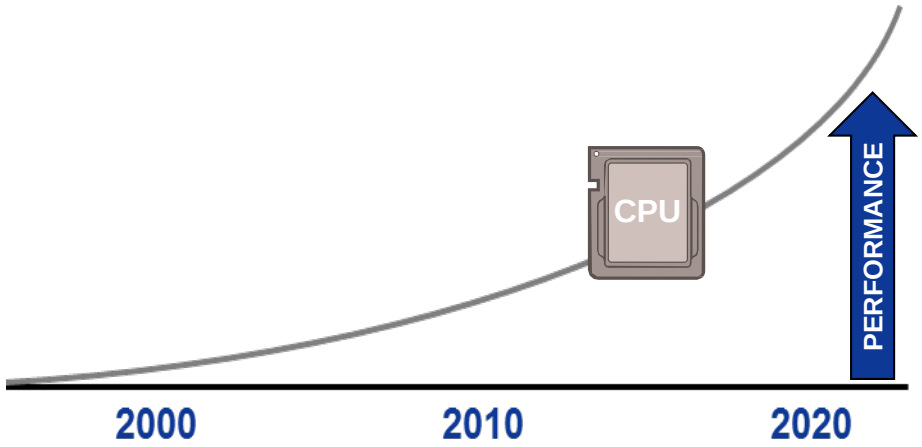
Flash in the Memory Channel

Maher Amer

CTO, Diablo Technologies

The Need To Feed

Processor performance continues to improve, creating massive opportunities for application acceleration…

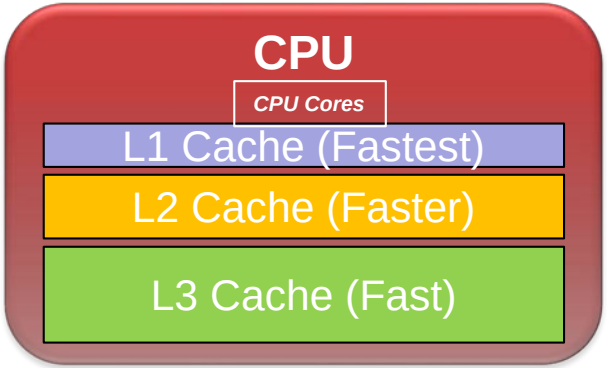


…but, to maximize their potential, those **hungry CPUs need to be constantly fed**…

…requiring **fast and consistent access** to **large amounts of data**

The Closer The Better

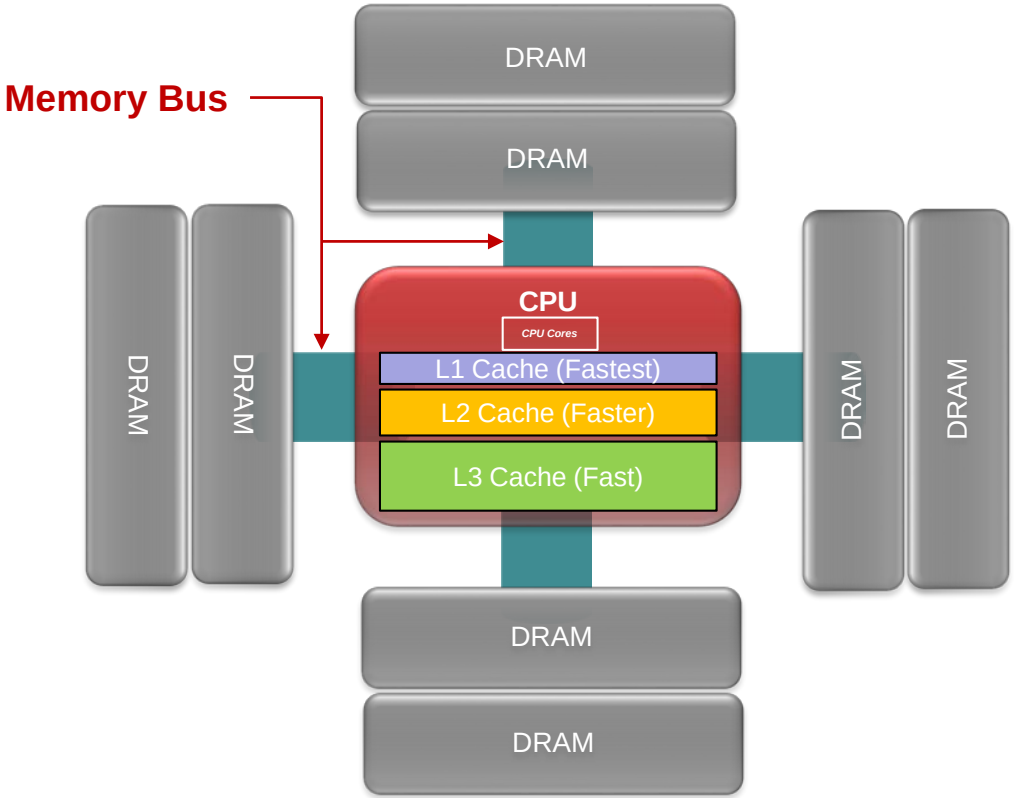
- **Data locality is crucial to system performance**
- On-die CPU caches provide the fastest data access…



- …but cache size is limited by silicon area

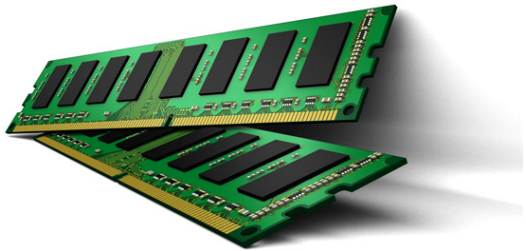
System Memory Architecture

- System Memory extends the “near CPU” data access domain
- Connectivity provided by the high-performance Memory Bus



A Good Solution…But With Tradeoffs

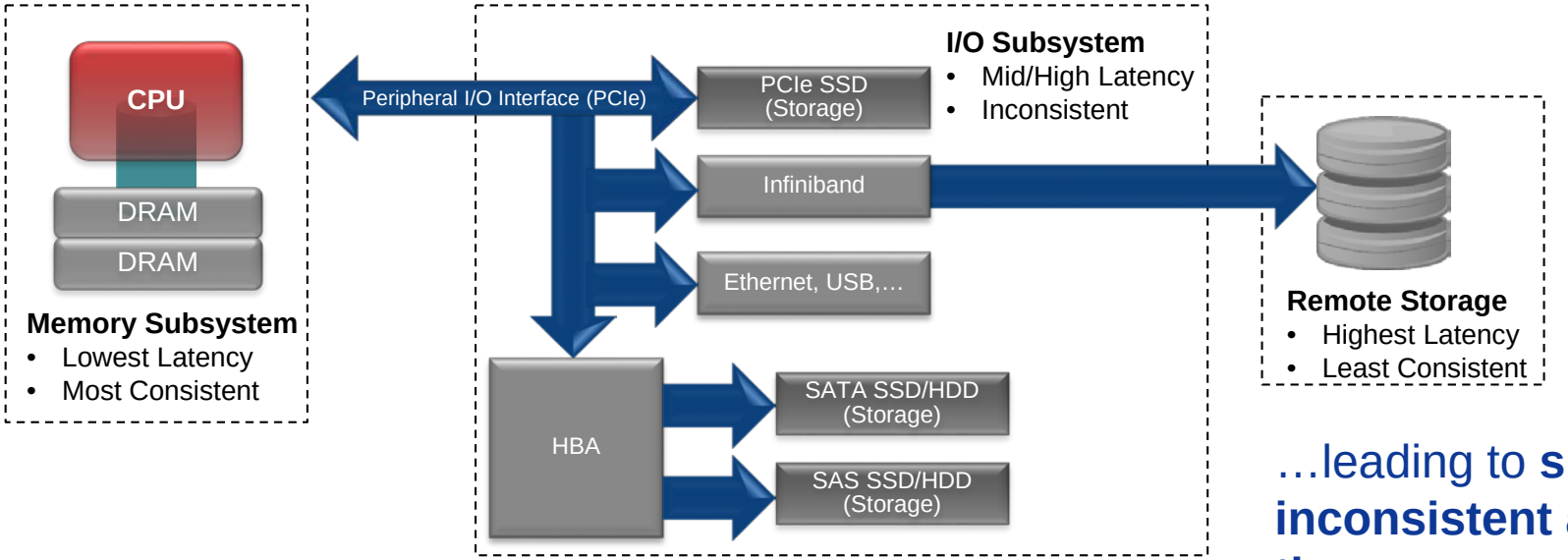
- The “hottest” application data resides in system memory…



- …but DRAM’s cost and capacity limitations force tradeoffs

Storage = A Necessary Evil

Historically, **crucial data** has been forced to reside **outside the memory subsystem**…



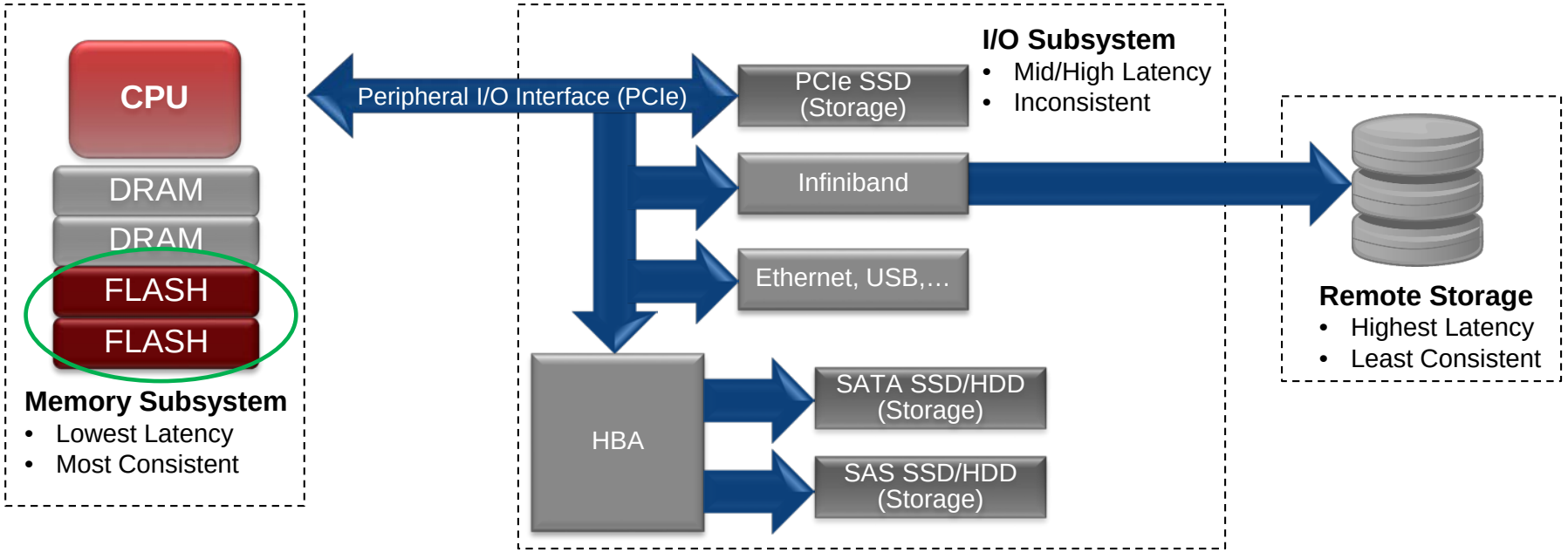
…leading to **slow, inconsistent access times**

“How can we improve?”

The Solution

- Expose the cost and capacity advantages of non-volatile memory (e.g. NAND Flash)…
- …on the highest-performing, most parallel bus in the system…
- …keeping more data local to hungry CPUs

Keep More Data In The Memory Subsystem

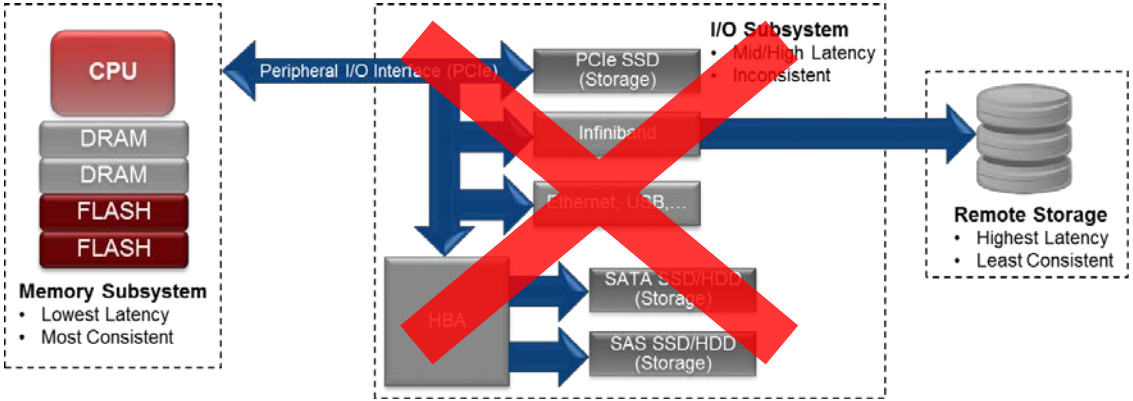


Data Remains Close To CPU

Parallel Access = Performance and Consistency

Flash On The Memory Bus…As **Storage**

- Eliminates Storage Bottlenecks
  - Bypasses I/O Subsystem…Shortens Path To CPU
  - Eliminates Contention With I/O Devices
  - Memory Bus Provides Massive Parallelism

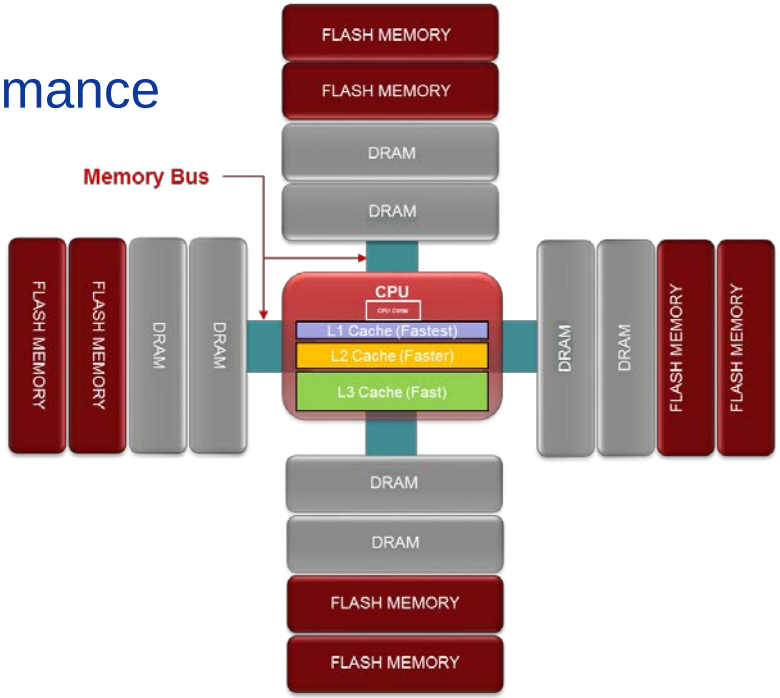


Flash On The Memory Bus…As **System Memory**

- Expands System Memory Capacity
  - Flash Density Enables More Memory Per Socket And Per Server
  - Improves Performance
    - Processes Larger Data Sets In-Memory
    - Minimizes Hops Across QPI And Network
  - Minimizes Infrastructure Costs
    - Leverages Economic Advantages Of Flash
    - Minimizes Number of Servers Per Deployment

Summary and close

- Data locality is crucial to application performance
- Keeping more data in-memory is solution
- Flash On The Memory Bus is the enabler
  - Leverage existing memory architecture
  - Huge performance and efficiency benefits



THANK YOU