Augmenting Memory Capabilities for Exascale Systems

DOE FastForward Research Project
PIs: Ravi Nair, Jaime Moreno, Douglas Joseph
IBM TJ Watson Research Center

© 2014 IBM Corporation

Memory subsystem will contribute an increasing fraction to power, reliability and cost of supercomputer systems

- **DRAM scaling has slowed faster than logic**
  - **More DRAM dies needed for each processor die to keep traditional compute to capacity and bandwidth ratios**
  - **Higher power, lower reliability, higher cost**
- **Memory has played limited role in the improvement in supercomputer performance**
  - **DRAM has not delivered 2x/year improvements in capacity (density) nor bandwidth**
  - **DRAM has become expensive relative to computation**
- **Alternative memory technologies not yet able to fill the emerging needs**
  - **Likely to be part of incremental changes throughout time**
- **Hybrid memory subsystems emerging, consisting of conventional DRAM and some low-cost dense alternative memory**
- **Memory performance should be included in the metric for evaluating supercomputers**

© 2014 IBM Corporation

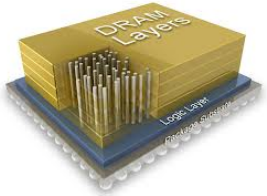
Novel memory devices, technologies and ways to exploit them are needed to maintain the 2x/year trend in supercomputers

- **The way how memory is partitioned and used by software was the source for the last significant paradigm shift in supercomputing**
  - **Large shared memory was replaced by message-passing between compute nodes with smaller local memory**
- **Data manipulation capabilities are needed in traditionally passive memory structures, beyond simply storing bits of data**
  - **Transform data closer to where it resides, before transporting it**
    - ⋆ **Outer level caches, SMP nest, memory controller, memory**
- **Advanced intra-memory management mechanisms are needed for improved efficiency, resilience and performance**
- **More intelligence is needed in data transportation media**
  - **Busses, switches, network**
- **Application developers may need to change algorithms**
  - **Use less memory per compute unit**
  - **Enhance computation locality at the memory level**

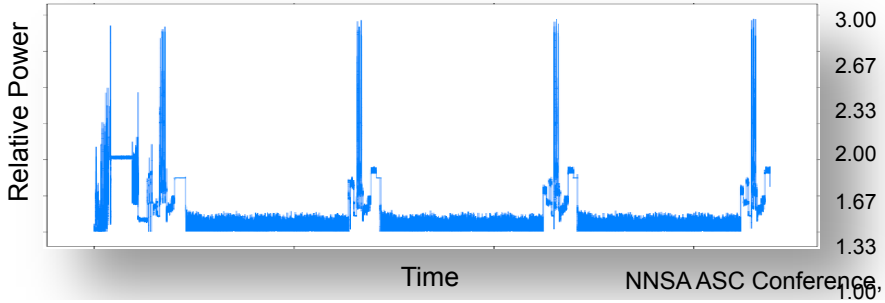
© 2014 IBM Corporation

Power/Thermal/Reliability of future memory devices, such as next generation Hybrid Memory Cube (HMC), are being explored in detail

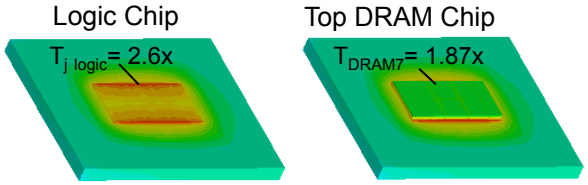
- Simulation Infrastructure
  - Benchmark based memory traces from a processor+memory simulator
  - Cycle approximate SystemC simulator models internal operation of HMC
  - Detailed HMC operation trace provided to power model

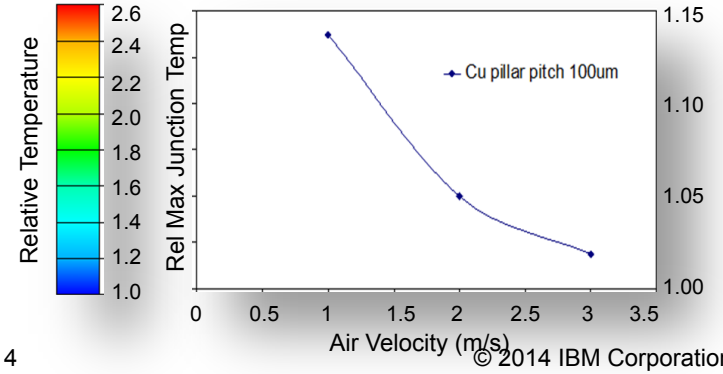


- Benchmark Driven Power Modeling
  - Power model provides component-wise utilization dependent power analysis
    - Power computation based on RTL power model and HMC operation trace
    - Support for power analysis of link gating and open/closed page policy



- Reliability Modeling
  - Study reliability impact of ECC codes, voltage, latches, etc.
- Thermal Modeling
  - Explore ways to improve thermal characteristics
  - System-level model details laminate substrate, Cu lid, TIM, heatsink, air flow
  - Micro-level model details materials and geometry at chip interfaces (thickness, dielectric material, metal loading, size/pitch/ material of micro-pillars)





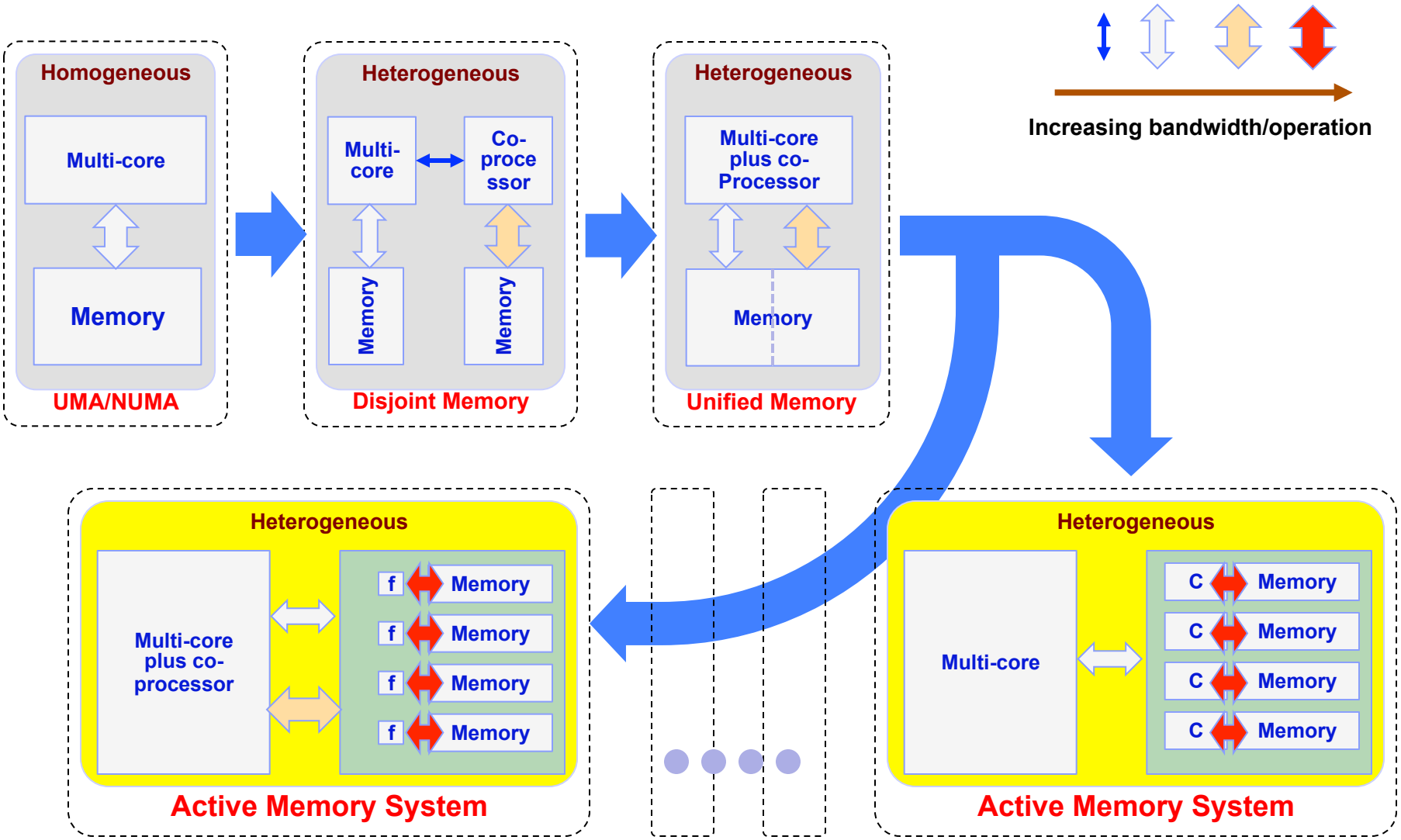
© 2014 IBM Corporation

Systems and programs must evolve towards an execution model that minimizes the movement of data to/from memory

- **Data movement consumes a large fraction of energy in a computation**
  - **Integration and packaging technologies help but only to some extent**
- **Larger gains are possible from consciously reducing data movement**
  - **Evolution towards "data movement aware" architectures and applications**
- **Exploitation of multiple forms of memory and their unique attributes will force a new significant paradigm shift**
  - **Programming models and applications must be developed with awareness of the type of memory where data reside, and the associated capabilities**
  - **Non-volatile nature of some types of memory present opportunities and restrictions**
    - ⋆ **For example, better error tolerance and lower power, but longer write latency and lower endurance**
  - **Similar issues arise even for same memory technology but different device characteristics, such as density, latency, and cost**
  - **Memory bandwidth is usually a differentiator that needs management by the application or system software**

© 2014 IBM Corporation

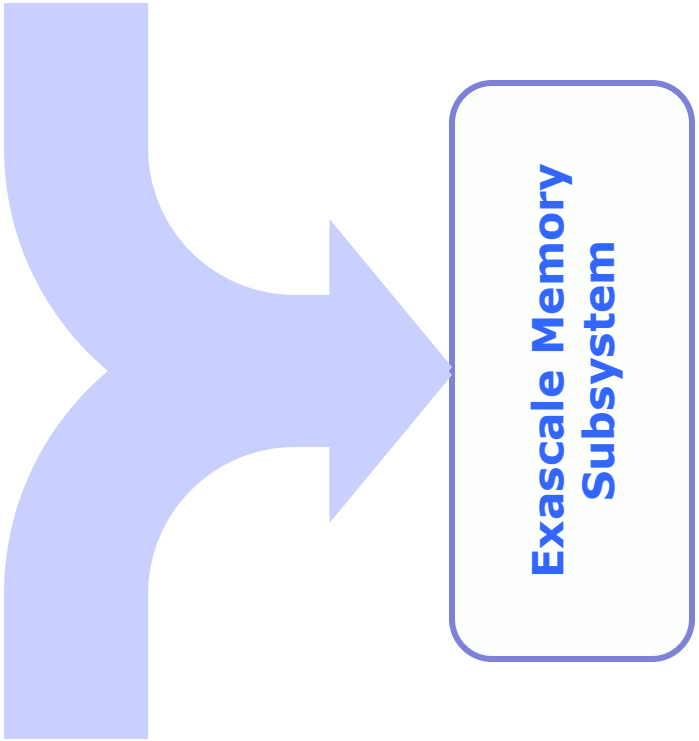
Node memory architecture has been and will continue to evolve in response to system needs



© 2014 IBM Corporation

Research and development thrusts are needed to cover the entire hardware/software stack for Active Memory Systems

Programming features	**Data locality representation** **Programming interface** **…**
Concurrency features	**Concurrency enhancements** **Synchronization** **Coherence** **…**
Data movement features	**Locality enhancements** **Data placement and transfer** **Granularity of transfers** **…**
Architecture features	**Functionality close to memory** **Power efficiency** **Reliability** **…**
Device improvements	**Power efficiency** **Reliability enhancements** **…**
Emerging technologies	**Dense memory devices** **Storage-class memory (non-volatile)** **…**



© 2014 IBM Corporation

Active Memory Systems have significant implications and impose new requirements on models and applications

- **Exascale systems will be workflow driven, with data located at different levels of the memory and storage hierarchy**
  - **New and challenging problems for software and application developers**

- **System view must encompass the whole storage hierarchy within and across nodes**
  - **Computation essentially a series of transformations on data**
  - **Orchestration of data manipulation within and across all levels of the system**
  - **Mechanisms for reasoning about this view of data and computation**
  - **New software APIs, new programming abstractions**

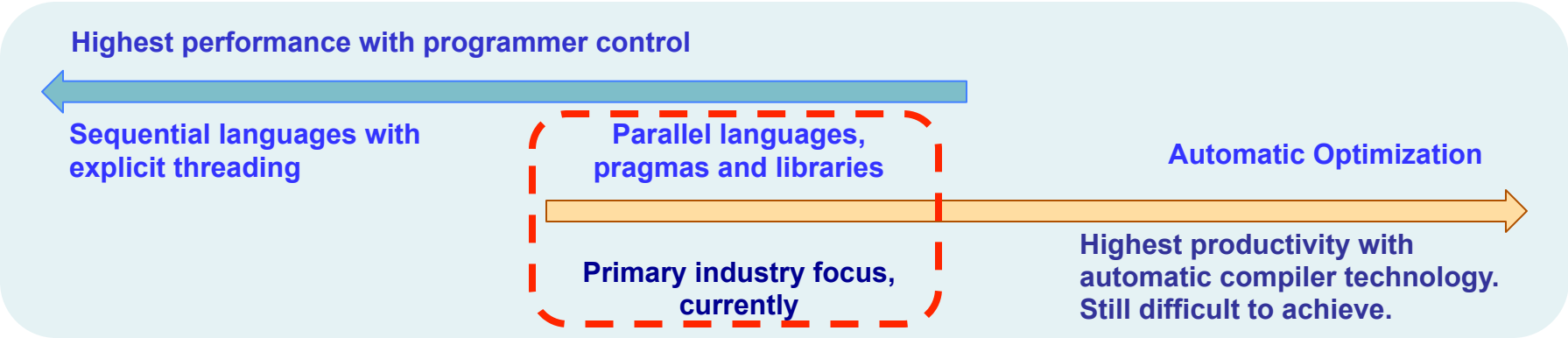
- **Current models are primarily focused on processing**
  - **Computation and control are the primary drivers**
  - **Data placement and communication are secondary considerations**
  - **Models need to be extended to support "data movement aware" principles**

- **Models must evolve to encompass all aspects of data management and computation, requiring new abstractions**
  - **Compute, data and communication are equal partners**
  - **Outer level abstraction providing means to reason about data and associated attributes**
  - **New/higher and lower level models co-exist**

© 2014 IBM Corporation

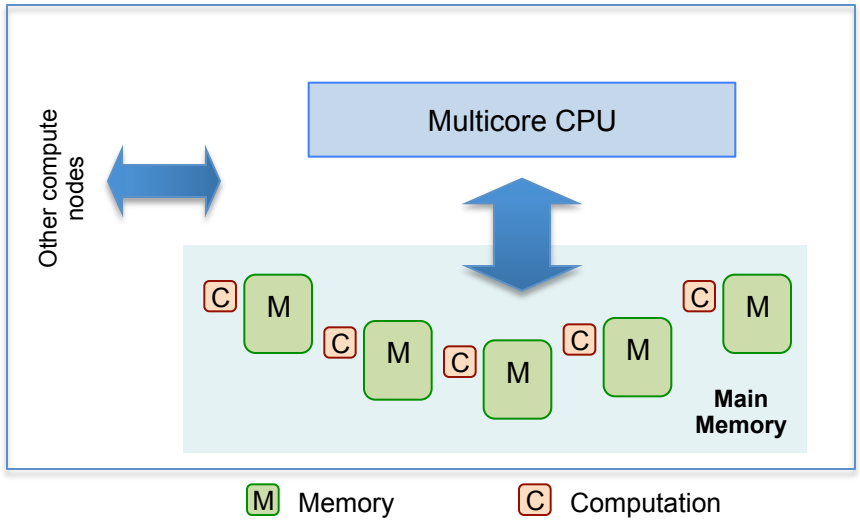
Expectations on programming models and tools for near-data processing have a broad and sometimes conflicting nature

- **Constraints**
  - Very large installed code base, very large number of kernels (loops)
  - Portability, incremental transition for existing code, extensible, predictable, etc.
- **Programming model needs**
  - Able to manage the node memory hierarchy, enhancing locality and minimizing data movement
  - Memory allocation model, with abstractions for memory organization and capabilities
  - Partitioning of data and associated computation within a node
- **Potential programming model approaches**
  - New or emerging languages, enabling revolutionary alternatives
  - Directives-based (e.g., OpenMP)
  - Standardized extensions to existing languages



© 2014 IBM Corporation

IBM FastForward project focuses on breakthrough innovations on Active Memory System Architecture and programming models



Novel Ideas

Novel forms of computation near memory.
Reduction in data movement and associated overhead.
Advances in programming models, compiler, operating system and runtime environment.
Novel leverage of emerging memory technologies.
Advances in memory efficiency.
Advances in memory system integration, power and reliability management.

Impact

Large reduction of data movement.
Significant improvement in system level performance, power efficiency, and reliability.
Successful exploitation of novel architecture features while abstracting the hardware complexity, enabled by evolutionary and revolutionary approaches.

Principal Investigator(s): Ravi Nair, Jaime Moreno, Douglas Joseph, IBM TJ Watson Research Center

Milestones/Dates

- **Architecture definition and assessment, programming model and software environment specification, application analysis and evaluation.**
  Sep 2014

© 2014 IBM Corporation