SAND2014-4109P

Exceptional service in the national interest

Sandia National Laboratories



Abstract Machine Models

S.D. Hammond and Computer Architecture Laboratory Team (SNL and LBNL)
Scalable Computer Architectures, Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY NNSA

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Introduction

Sandia National Laboratories

- Concern that categories/types of hardware architecture are proliferating
  - Disparate types include GPUs, multi-core CPUs, integrated GPUs, specialized accelerators and many-core CPUs
- Adding to the complexity for designing applications to run at Exascale
- Think how easy MPI has been as a "model" of hardware
  - Serial control flow punctuated by explicit exchanges of data
  - Specifics of the underlying hardware are abstracted away
- Time for a new (hardware?) model for Exascale?

What is CAL?

COMPUTER ARCHITECTURE LABORATORY

EXASCALE DESIGN SPACE EXPLORATION

- Facilitate hardware simulation, codesign and modeling across the Codesign centers
- Bring vendors, projects, programming models etc together
- Act as a collaboration point
- Based at LBNL (DOE/ASCR) and Sandia (DOE/NNSA/ASC)
- POC: John Shalf (LBNL) and Jim Ang (SNL/NM)

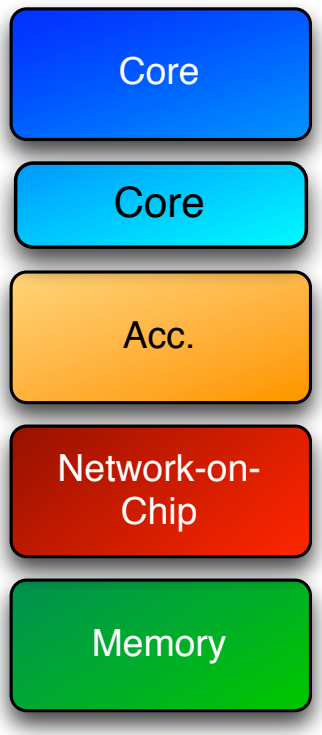
Short Summary of This Talk

- Motivations from hardware
- Bottom Up Approach
  - Homogeneous Multicore Processor Model
  - Homogeneous Multicore Processor Model + Discrete Accelerators
  - Integrated Multicore Processor and Accelerator Model
  - Heterogeneous Multicore Processor Model
- Top Down Approach
  - Call to “build the processor of for **YOUR** application”
  - Think of this as a System-on-chip (just for you)

Motivations in Hardware

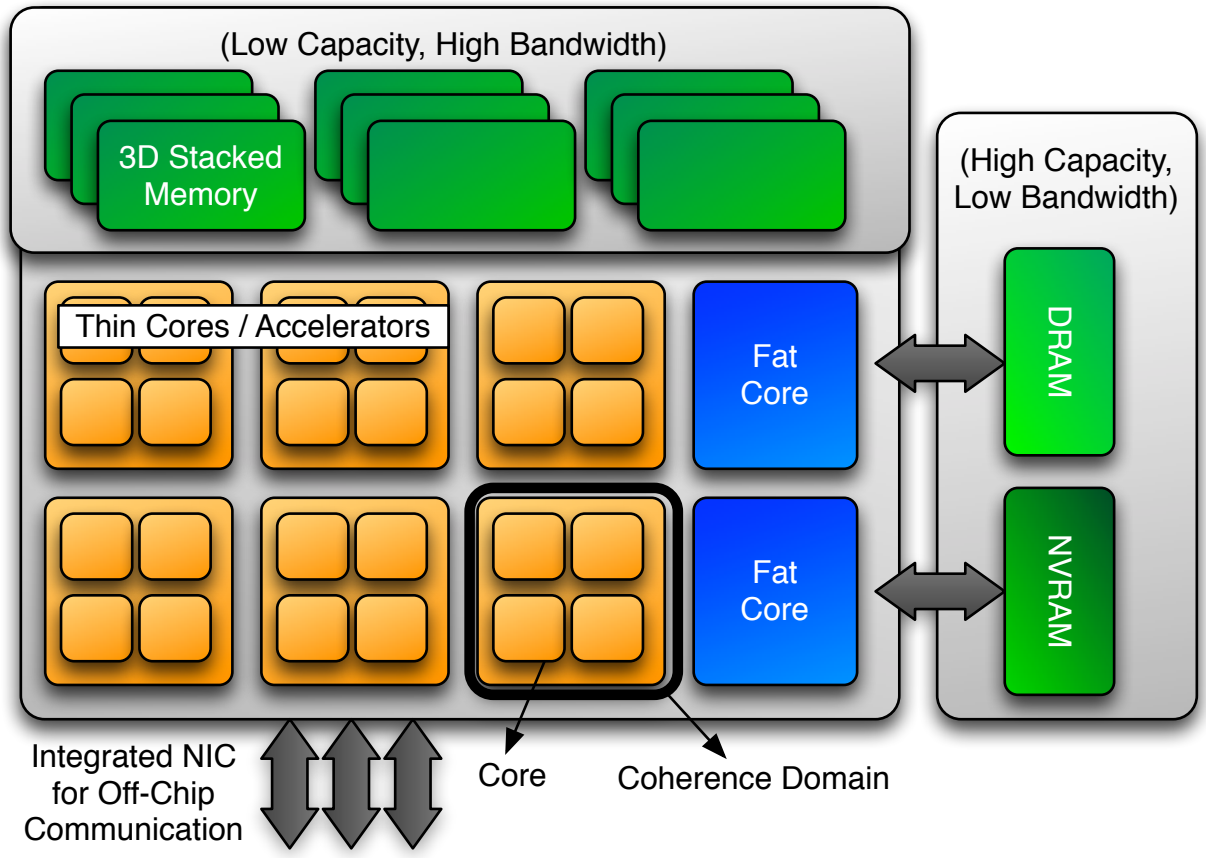
- Increasing levels of Parallelism
- Variegation of "processing element" types
  - See growing use of *thin* cores for wide-parallel execution
  - Continued use of *fat* cores for serial execution (or limited parallelism)
- Limited coherency domains
- Multiple levels of (explicit) memory
- Sophistication of on-chip networks to deliver performance

Building Blocks of a Hardware Model



- Cores - *medium* and *fat* cores
  - Standard serial processors
- Accelerators - *thin* cores
  - Think highly threaded and/or wide (>32) vector
- Network on chip
  - Something more sophisticated than a ring
- Memory
  - Multiple levels - HMC/HBM/WideIO, DRAM and NVRAM

Overarching Model

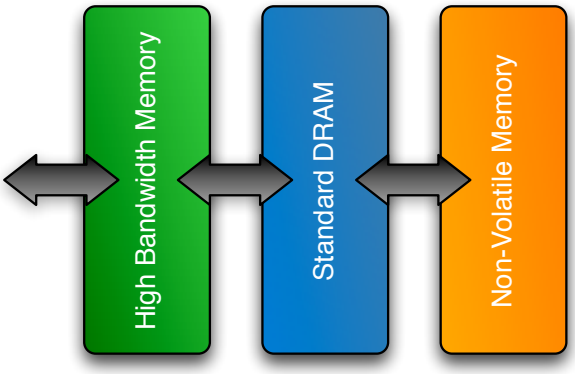


Hardware Reality

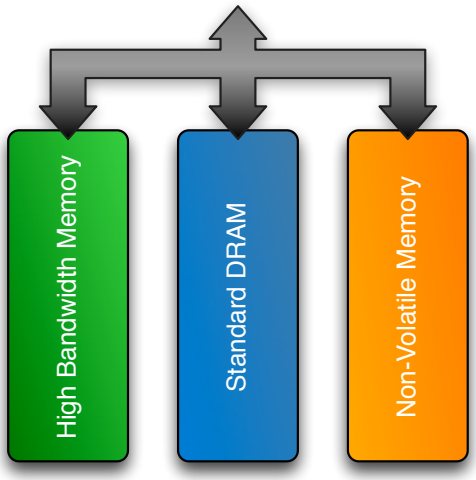
Sandia National Laboratories

- Many vendors want to pursue processors for *existing* markets
  - Bigger than HPC?
- Probably won't see *all* of these features in a single die
- Leads us to think about *plausible* models for the future
  - Economically possible, performance possible models which are more likely to be delivered
- Subject to *perfecting* by codesign
  - We *can* change the future but we may not be able to radically redesign every aspect of the processor

Models of Memory



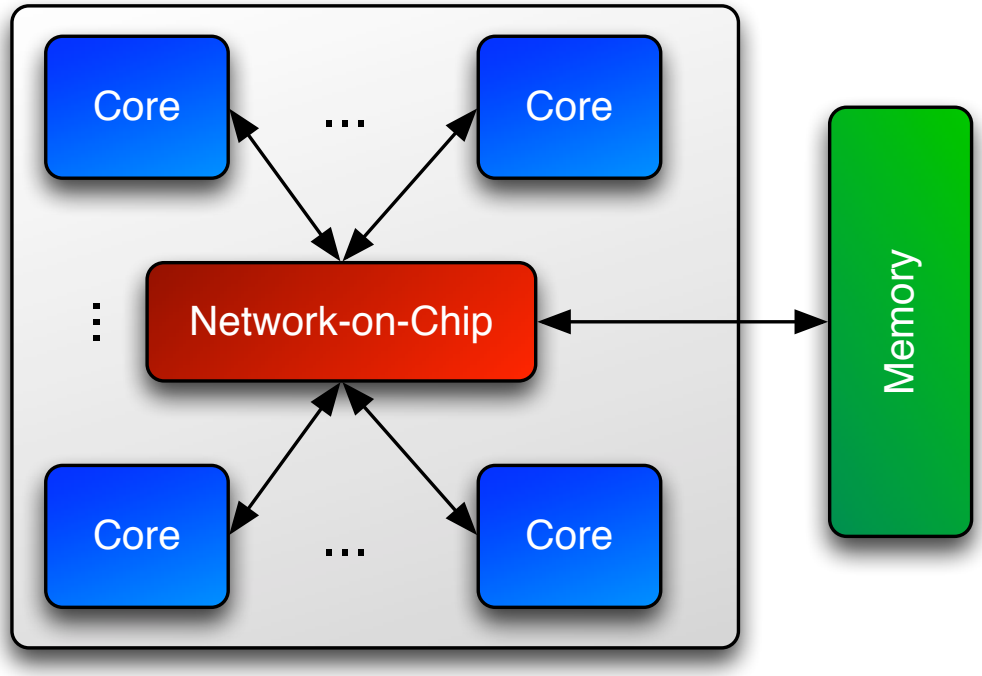
Cache Model



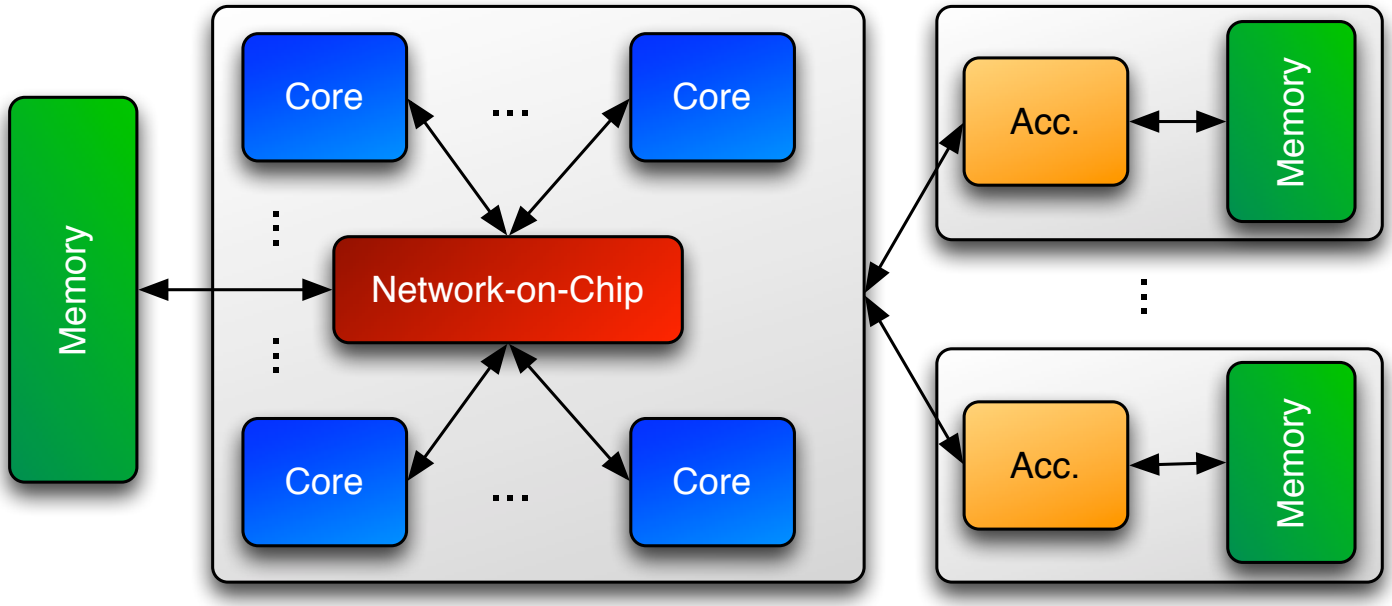
Explicit Allocation Model

- Levels may be used as a cache
  - Bad option for graphs, analytic applications, good if you *can* stream
- Explicit allocation ("partitioned hardware address")
  - Headache of the programmer but performance will be high
- Combination of the above?

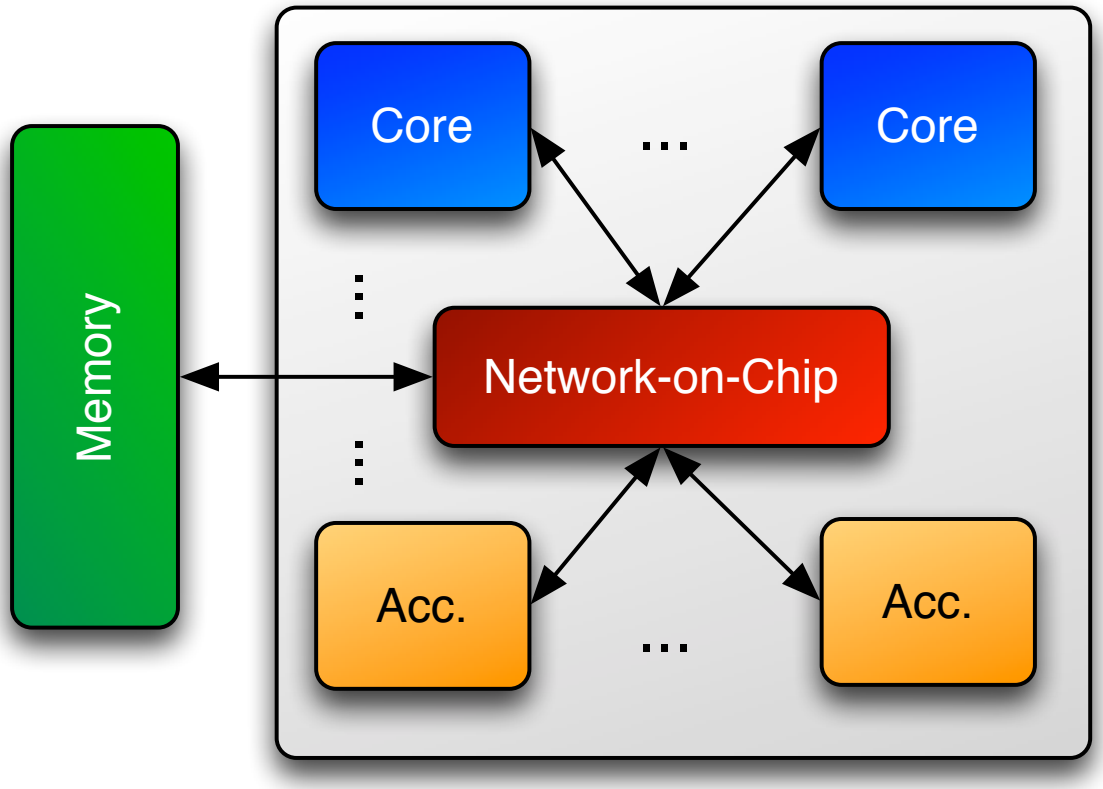
Homogeneous Multicore



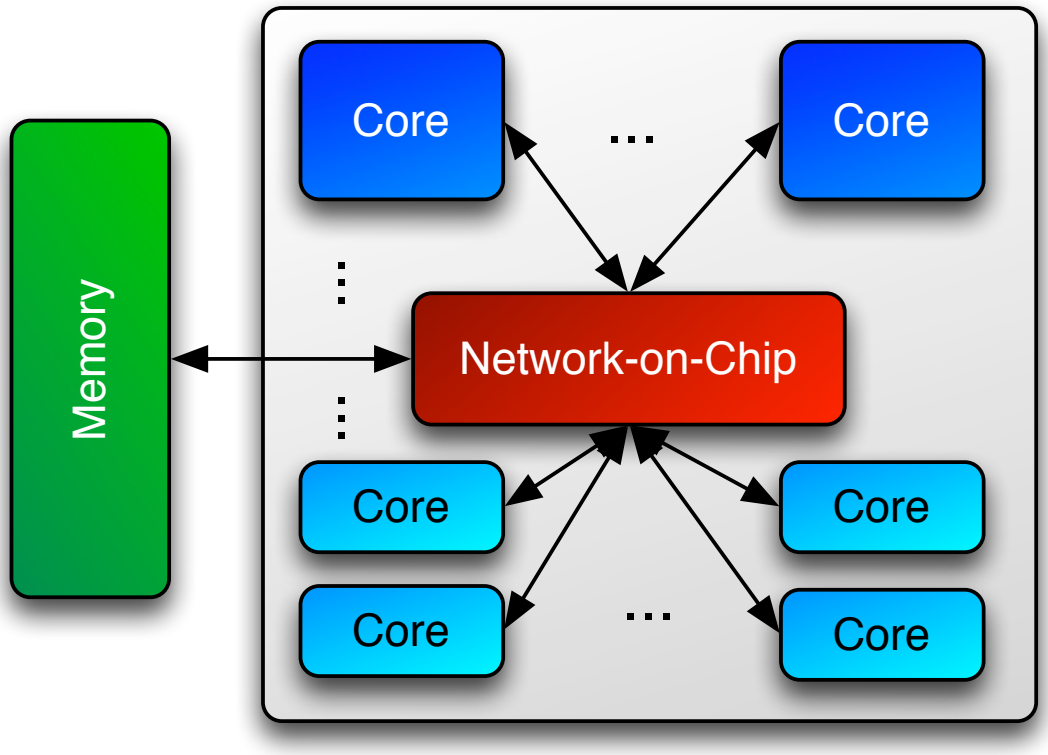
Homogeneous + Discrete Accelerator



Integrated Accelerator



Heterogeneous Multi/Many-core

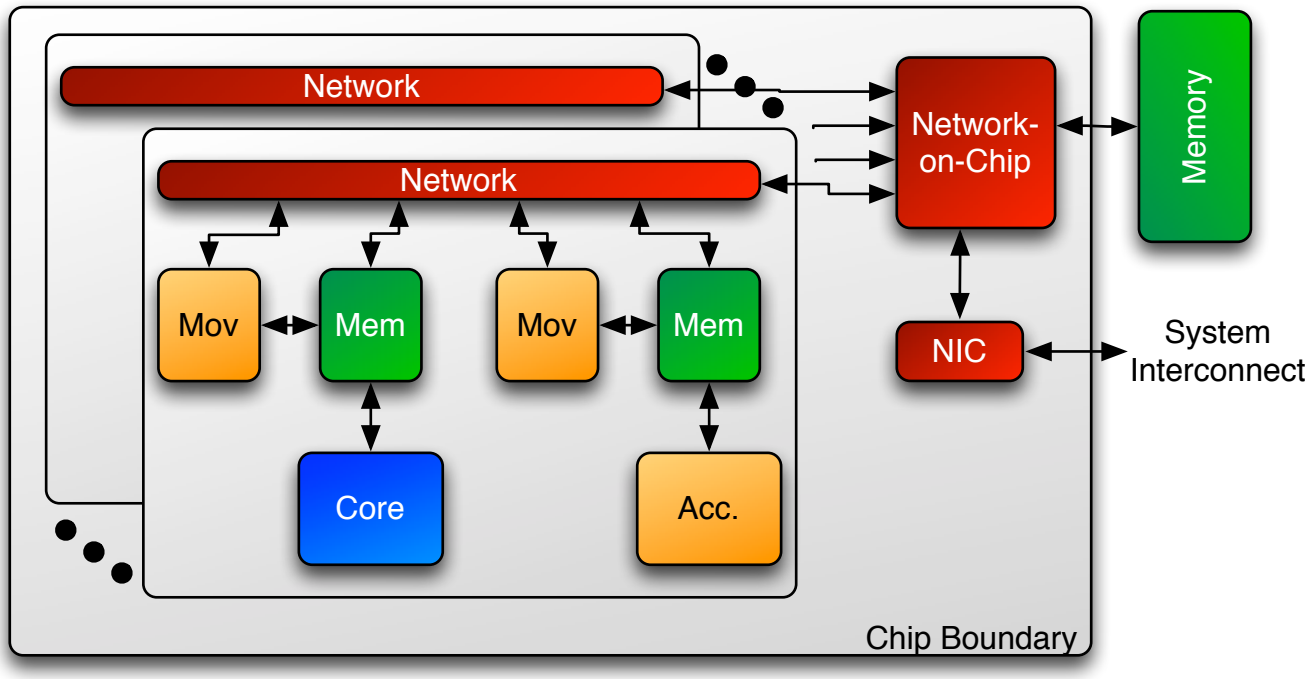


Top Down Approach

Sandia National Laboratories

- What if you *could* pick and choose features?
- Possibility if the Exascale Computing Initiative (ECI) is funded by congress
- Start with existing IP blocks but cherry pick what *you* need
  - Potential for greater overall performance
  - Potential to save huge cost (new IP costs a **lot** of money, SoC is cheap(er))
- Many vendors will offer some IP blending/selection capability

Concept Design



Courtesy of Dave Resnick (Sandia National Laboratories)

Questions for You

Sandia National Laboratories

- Do abstract machine models help you?
- Can you see parts of your algorithm mapping onto the hardware?
- What extra things do you think we need?

Proxy Architectures

Sandia National Laboratories

- Proxy architectures are an Abstract Machine Model with parameters supplied
  - i.e. we add detail to the model
- See the report for details
  - Does this help you map your algorithm?
- What *additional* features do we need?
- Assume between 4 and 16 TFLOP/s nodes
  - 20,000 - 60,000 machine nodes?

Sandia
National
Laboratories

Exceptional service in the national interest