Moving *Forward: Pathfinding to Exascale

Smoky Mountains Conference 2015

Larry Kaplan
Chief Software Architect
Agenda and Acknowledgment

- DesignForward
- FastForward2
- DesignForward2

This work is funded in part by the US Dept of Energy
Broader Cray Efforts

● **Programming environment**
  ● Focus on programmability and usability
  ● Evolution of MPI + OpenMP
  ● Revolution via Chapel

● **System Software**
  ● Memory and storage hierarchy
  ● System management
    ● Energy, RAS, etc.
  ● Workflow support
DesignForward Project Overview
Cray’s DesignForward Goals

- Protocols and APIs for an HPC capable open network
- Provide Cray and DOE with a path for using and managing commodity-based networks for HPC
- Recommend low level network API
- Recommend network management system
DesignForward – Network API

Goal: Specification of a proposed network API (or extensions to an existing API), including potential specification of a plug-able architecture, for implementation on multiple network types

- Focused on support for MPI applications
- Other programming models also considered
- Facilitate portability between different networks
DesignForward – Network API Status

- Reviewed existing and recently proposed APIs and their pros and cons
- Consulted with community runtime/model efforts
- Focused on OFI/libfabric
  - Open source model with dual license (BSD and GPLv2)
  - Framework and base interfaces for portability
  - Provider extensions allow for vendor customization
- Working with LANL on a prototype GNI provider layer for existing Cray XC systems
- Also evaluating libfabric clients including MPI and Chapel
OFI/libfabric Structure

Fabric Interfaces
- Control Interface
- Message Queue
- Addressing Services
- RMA
- Tag Matching
- Atomics
- Triggered Operations

Fabric Provider Implementation
- Control Interface
- Message Queue
- Addressing Services
- RMA
- Tag Matching
- Atomics
- Triggered Operations

Frame work defines multiple interfaces

Vendors provide optimized implementations

http://ofiwg.github.io/libfabric

Copyright 2015 Cray Inc.
Goal: Describe proposed multi-level network management architecture

- Considered important HPC use cases and actors
  - Partitioning, bandwidth management, diagnosis (func and perf)
  - Developer, User, Site and System Administration, Service, Network Designer, Cloud/Self Service

- Specified network management API requirements
  - Desire modularity and support for 3rd party integration

- Reviewed existing network management tools
  - None met all of the requirements
DesignForward – Network Management (2)

- Proposed a new multi-level network management architecture capable of managing multiple disparate networks
- Prototyped subset of functionality in open network management system
- Hybrid approach with integration support
  - Relies on components from OpenStack and other open source
  - Proprietary components also supported
FastForward 2 Project Overview
About Cray’s FastForward 2

Goals:

● Research and develop the technologies necessary for ARM ISA-based processors to be competitive in HPC
  ● Both hardware and software components
  ● Performance and Performance per Watt

● Research and specify the design of Exascale (~2020 & ~2022) ARM ISA-based nodes, including:
  ● SoC designs
    ● Core and un-Core
  ● Memory architectures and interfaces
  ● Network interfaces
  ● Programming model
ARM ISA HPC Improvements

- Evaluate current ISA and its suitability for HPC
- Consider enhancements to the ARM ISA
- Support for simulation
- Support for data analytics
Core Architecture Study

● Performance, Power, and RAS all critical

● Single-Thread vs. SIMD vs. Thread-Optimized
  ● “Task Cores” with different optimization goals
  ● Re-configurable?

● ARM TAU
  ● Throughput optimized cores with ARM ISA
SoC Architecture Study

- Homogeneous vs. Heterogeneous
  - But single-ISA

- Core mixes and configurability
  - How to provide the best efficiency?

- On-chip network
  - Synchronization

- Memory Architecture
  - Will explore future technologies with memory vendors
  - Goal to have *commodity products* able to serve the HPC market
FF2 NIC Overview

- DOE requirement for a very high rate of small messages, e.g.
  - 500M/s × 256B MPI
  - 2B/s × 8B Put/Get

- FF2 concept of meeting these requirements using multiple highly integrated NICs per socket

- Focus on SoC integration of NIC to support these rates
Simulation Plans

● **ARM: Flexible open exploration tools (GEM5)**
  ● Easy exploration of wide parameter spaces and ISA
  ● Integration into open system models (SST)
  ● Available for wide use (customers/ecosystem)

● **Broadcom: Detailed implementation models**
  ● Accurate simulation of potential implementation decisions
  ● Power/Energy modeling

● **Cray: High-Speed applications analysis**
  ● High-throughput simulation/modeling tool (SAGE) allows more realistic app sizes to be examined (also good for trace generation)
  ● Capable of full-socket (many-core SoC) simulations at reasonable rate

● **DOE: Mini-apps**
  ● Using these apps for most simulation work
  ● Will leverage “Hackathons” to engage DOE in evaluation of technical findings
DesignForward 2 Project Overview
DesignForward 2

● Exascale system study
● Components
  ● Nodes
    ● Evaluate primary candidate architectures
  ● Networks
    ● Injection and global bandwidth requirements
    ● Network features
  ● I/O
    ● Hierarchy
    ● Network interaction
Key Trends in Processor Design

- **Some technique for high flops/W**
  - Vectors
  - Tightly coupled GPUs

- **Aggressive memory hierarchy**
  - On-package HBM stacks
  - Larger memory off package (likely NVM)
  - Conventional “DDR” DRAM may eventually disappear
    - How much capacity is required in “near” memory, given the likely large bandwidth cliff to off-package memory?

- **Three architectures emerging**
  - Homogeneous Many-core architecture
  - Multi-core CPU with accelerator architecture
  - Heterogeneous Multi-core architecture
Node Architectures to be Examined

DesignForward 2 (continued)

● **Execution models**
  ● Mapping to target architectures
  ● High-level runtime specification
  ● Focus on MPI/OpenMP and Chapel

● **System organization**
  ● Impact on performance, resilience and reliability, energy use
  ● Hardware and software features for productivity

● **Application mapping**
  ● Map DoE proxy apps to programming and execution models
  ● Consider impact on performance and efficiency
Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Other names and brands may be claimed as the property of others. Other product and service names mentioned herein are the trademarks of their respective owners.

Copyright 2015 Cray Inc.