NSTX-U Advances in Real-time Internode Communications

Keith Erickson\textsuperscript{1}, M. Dan Boyer\textsuperscript{1}, and David Higgins\textsuperscript{2}

11th IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research
Greifswald, Germany
May 10, 2017
Outline

• NSTX-U Overview
• NSTX-U Real-time System
• Computing Architecture and Operating Environment
• Expanding the RT System
• Reflective memory technology
• Empirical testing and results
• Summary
NSTX-U Mission Elements:

• Explore unique ST parameter regimes to advance predictive capability - for ITER and beyond

• Develop solutions for plasma-material interface (PMI)

• Advance ST as Fusion Nuclear Science Facility and Pilot Plant
NSTX-U will access new physics with 2 major new tools:

**New Central Magnet**

Higher $T$, low $\nu^*$ from low to high $\beta$  
→ Unique regime, study new transport and stability physics

**Tangential 2nd Neutral Beam**

Full non-inductive current drive  
→ Not demonstrated in ST at high-$\beta_T$  
Essential for any future steady-state ST
NSTX-U will have major boost in performance

New Central Magnet

- 2× toroidal field (0.5 → 1T)
- 2× plasma current (1 → 2MA)
- 5× longer pulse (1 → 5s)

Tangential 2nd Neutral Beam

- 2× heating power (5 → 10MW)
  - Tangential NBI → 2× current drive efficiency
- 4× divertor heat flux (→ ITER levels)
- Up to 10× higher nT\(\tau_E\) (~MJ plasmas)

Performance goals:
NSTX-U had scientifically productive 1st run year

- Achieved H-mode on 8th day of 10 weeks of operation
- Surpassed magnetic field and pulse-duration of NSTX
- Matched best NSTX H-mode performance at ~1MA
- Identified and corrected dominant error fields
- Commissioned all magnetic and kinetic profile diagnostics
- New 2nd NBI suppresses Global Alfven Eigenmodes (GAE)
- Implemented techniques for controlled plasma shut down, disruption detection, commissioned new tools for mitigation
- 2016 run ended prematurely due to fault in divertor PF coil
  - Coil + other issues → major reviews of design, fab, procedures
• Analog and digital inputs and outputs cycle synchronously at 5 kHz
• VITA 17(.1) FPDP standard for parallel (serial) communication
• Serial connection allows ~1km distance between the computer and the devices
• Currently ~500 sensors and ~50 actuators
RT Computer Specifications

• Concurrent RedHawk 6.5
• Supermicro H8QGL
  – Opteron 6386 SE 2.8GHz
  – 4x16 cores = 64 cores
  – 64 GB Registered ECC DDR3
• 6 PCIe slots in two banks
  – Serial FPDP I/O (SL-240)
  – Realtime Clock and Interrupt Module (RCIM)
  – CUDA capable video card
  – Dolphin PXH-830
Real-time Computer Architecture

- AMD Opteron 6386 SE
- Kingston Registered ECC Server Memory
- LSI2008 SAS RAID
- Seagate ST1000NM0063 with FIPS 140-2, 1.5Mh MTBF
- CUDA capable GPU
- Concurrent iHawk design
- Concurrent RedHawk Operating System (6.3)
- NightStar development environment
- Realtime Clock and Interrupt Module
Outsourcing OS Expertise Provides Cost-Effective Development Strategy for NSTX-U

• Concurrent Computer Corp. offers RedHawk
  – Based on RedHat (or CentOS)
  – Custom kernel to support deterministic run time behavior
  – NightStar analysis package permits performance optimization

• Concurrent provides custom drivers for COTS hardware
  – Full software support and Return-to-Factory warranty
  – Source code available

• Guaranteed process dispatch latency of less than 10 us
• Removes the need for in-house kernel hacks
• Concurrent support has proven invaluable for NSTX-U
Many-core RT system with development twin served NSTX well for over a decade

1. Multiple technological evolutions
   - SkyBolt I ➤ SkyBolt II ➤ SunFire v40z ➤ iHawk

2. Adding Protection to the RT Control computer has significantly reduced available cores

3. Physics demands require a flexible configuration with more cores at higher clocks
NSTX-U RT CPU Usage

PROTECTION SYSTEM
- 18 Algorithm cores
- 3 I/O cores
- 4 Monitoring cores

CONTROL SYSTEM
- 3 I/O cores
- 4 General Use cores
- 16 rtEFIT “Slow” cores

OPERATING ENVIRONMENT
- 16 User cores (shared)
- 16 OS cores (shared)

25 cores
23 cores
16 cores

64 cores total
### RT Computing Expansion Scenarios

**Option 1: New AMD System**
- AMD left the server market in 2012 following Opteron 6300
- New Zen-based servers not released: Specifications? Reliability?
- No motherboard vendor support yet

**Option 2: New Intel System**
- More cores, more instructions per clock
- Technology curve lacking: slower clock/core, lackluster gain since 2012
- Cost-prohibitive: E7-8894 is ~$10k per CPU

**Option 3: Dolphin Interconnect**
- High performance, *deterministic* interface to additional nodes
- Allows continued use of existing assets with scalable growth
- High cost of maintaining twin development systems
Dolphin Interconnect Solutions Supports Multiple Possibilities

- Founded 1992 in Oslo, Norway
- Long history of PCI and PCIe experience
- Product lines used in military aircraft
  - US: F-35
  - France: Rafale
- Products in use in fusion (WEST, etc.)
- Strategic partners with Concurrent for RT deployment
PXH-830 Gen3 PCIe I/O

- Non-transparent bridging host adapter connects two devices (host or switch) using PCIe protocols
- Gen 1/2/3 compatible, up to 16 lanes
- Switches use up to 8 lanes
- Theoretical performance characteristics
  - 32 Gbps per port, 4 ports = 128 Gbps (at Gen3 speeds)
  - 0.54us minimum latency (more on that later…)
  - Transmission distance
    - Copper: 9m
    - Fiber: 100m
NSTX-U Current Interconnection Layout

- NSTX-U Sensors and Actuators
- Development System
- FTC-1
- FTC-2
- FPDP1
- PCS-RT-3
- PCS-SRV-3
- PCS-RT-4
- PCS-SRV-4
- Ethernet Switch
- VLAN
- Warthog
- AT-2
- FPDP2
- A/D
- FPDP
- Ethernet
- Ethernet + IPMI
- Secondary Testing Rig

IAEA 11th TM on CODAC, NSTX-U Advances in Real-time Internode Communications, Erickson, May 2017
NSTX-U Future Interconnection Layout

- **NSTX-U Sensors and Actuators**
- **Proposed Expansion**
- **Dolphin Switch**
  - **PCS-RT-3**
  - **PCS-SRV-3**
- **PCS-RT-5**
- **PCS-RT-6**
- **PCS-RT-N**
- **Ethnernet Switch**
- **FPDP1**
- **Warthog**
- **AT-2**
- **FPDP2**
- **VLAN**
- **Development System**

Legend:
- A/D
- FPDP
- Dolphin
- Ethernet
- Ethernet + IPMI

IAEA 11th TM on CODAC, NSTX-U Advances in Real-time Internode Communications, Erickson, May 2017
Dolphin uses RDMA and Device Sharing

- **RDMA Transport mechanism bandwidth comparison**
  - Ethernet (iWARP, RoCE) ~40 Gbps, 10-25% overhead
  - OpenFabric / InfiniBand ~50-100 Gbps, 5-10% o/h
  - **Dolphin PCIe** ~130 Gbps, 1.5% o/h

- **No CPU, cache, or context switch overhead**
  - Remote Direct Memory Access (RDMA) transfers memory between computers without OS involvement
  - Device Sharing gives remote computers transparent access to PCIe devices

- **Pairing with RedHawk kernel limits jitter to < 2μs**
- **Latency scales linearly with data size**
Communication Path Bypasses CPU

Reflective Memory

[Diagram showing the bypassed communication path between CPU, DDR, Memory Controller, HyperTransport, SR-5690 Northbridge, GPU, PXH-830, and SL240.]
Communication Path Bypasses CPU

Device Sharing (Bridging)
Test Setup

Simulation runs mirroring current NSTX-U design

Simulation runs mirroring future NSTX-U design

- Biggest timing impact of various configurations is Gen2 vs Gen3 PCIe
- This is expected given direct connection between PCIe and memory
- Additional improved timing from copy modes vs size

Server

Client

Min Latency

1.7μs

1.3μs

0.5μs

Setup

- CPU Shielding
- RT Scheduler
- RT Priority
- Locked memory

Mini-SAS direct
Latency of Increasing Inter-host Reflective Memory Transfer Sizes

Worst Case: AMD Gen2 to AMD Gen2

About 1 μs jitter regardless of transfer size and latency

Latency scales linearly with data size
Newer Gen3 Systems Reduce Latency

Times shown indicate max round trip latency
Bars indicate min-max range

Intel Latency
AMD Latency

0 1 2 3 4 5 6 7
8192 4096 2048 1024 512 256 128 64 32 16 8 4 2 1 0

0 4 8 16 32 64 128 256 512 1024 2048 4096 8192


0 1 2 3 4 5 6 7
8192 4096 2048 1024 512 256 128 64 32 16 8 4 2 1 0

However, jitter is not as consistent (but it is inconsistently better, not worse).

<table>
<thead>
<tr>
<th>Size</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.982</td>
<td>2.222</td>
</tr>
<tr>
<td>4</td>
<td>1.000</td>
<td>1.374</td>
</tr>
<tr>
<td>8</td>
<td>0.988</td>
<td>1.376</td>
</tr>
<tr>
<td>16</td>
<td>0.992</td>
<td>1.417</td>
</tr>
<tr>
<td>32</td>
<td>1.013</td>
<td>1.434</td>
</tr>
<tr>
<td>64</td>
<td>1.014</td>
<td>1.995</td>
</tr>
<tr>
<td>128</td>
<td>1.030</td>
<td>1.386</td>
</tr>
<tr>
<td>256</td>
<td>1.061</td>
<td>1.438</td>
</tr>
<tr>
<td>512</td>
<td>1.132</td>
<td>1.594</td>
</tr>
<tr>
<td>1024</td>
<td>1.266</td>
<td>1.670</td>
</tr>
<tr>
<td>2048</td>
<td>1.530</td>
<td>1.971</td>
</tr>
<tr>
<td>4096</td>
<td>2.082</td>
<td>3.109</td>
</tr>
<tr>
<td>8192</td>
<td>3.174</td>
<td>4.316</td>
</tr>
</tbody>
</table>
Summary

- Real-time control is a critical component of the scientific mission of NSTX-U
- NSTX-U requirements are increasing faster than CPU technology
- New Dolphin interconnect has proven viable for deterministic, efficient, and scalable communication
  - Scalable deployment extends component life, removes cost of wholesale replacement
  - Scalable protocol allows linear timing vs payload size
- Biggest challenge: maintaining a complete duplicate is costly but highly valuable