Gnodal has solved the fundamental weaknesses in Ethernet

Family of 10/40Gb switches for the High Performance Data Center
Gnodal - Overview

- Founded March 2007
  - Vision: HPC network performance in Ethernet
- Expert core team
  - 20 year track record of high performance networks
  - Core team ex-Quadrics
- 14 patents filed to date
  - Focusing on anti-congestion
  - In-order packet delivery and load-balancing
  - Scalability
- Product Launch in May ‘11 with Shipments in June ‘11
  - Low-latency FSI
  - HPC Throughput
  - Medium-sized HPC
- Industry Leading Supply Chain
Ethernet on the Move

Multi-Pathing, DCB, Low-latency
Fundamental Problem

- Ethernet switches congest, leading to...
  - Reduced network bandwidth
  - High unpredictable latency
  - Limited scalability and over provisioning
  - Poor quality of service and lost business
  - Increased power consumption
  - Inefficiency and increased total cost of ownership

- Who cares: customer examples
  - Financial trading firms: lost trades
  - Internet search engines: slow data retrieval
  - Storage systems: lost data
  - Video: poor quality of service
  - Scientific simulations: performance and time
Product Capability

Gnodal’s System:
- Scales from 72 to 64000 ports
- Largest, most cost effective data center switch
- Highest per port bandwidth
- Best Performance, Power and Utilisation

Gnodal designed chip:
- **Solves fundamental congestion problem**
- Lowest network latency
- Lowest power consumption

---

Gnodal Peta
10/40Gb Ethernet ASIC
Industry Leading Partners

- Plexus Contract Manufacturer
  - Worldwide production
  - Networking specialist
    - Juniper and others

- IBM Microelectronics
  - WW Leader in Technology

- Arrow Electronics
  - ASIC and component supply chain

- Aricent Stack
  - Also partners with Unisys, Juniper, Avocent
Gnodal Fabric – Ethernet at the Edge!
GS Series Switches
<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS7200</td>
<td>Top of Rack Switch</td>
<td>72 x 10GbE SFP+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replaceable Unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5Tb/s Throughput</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hot-swap PSU / Fans</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1RU, 112 - 262W</td>
</tr>
<tr>
<td>GS4008</td>
<td>Top of Rack Edge Switch</td>
<td>40 x 10GbE SFP+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replaceable Unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5Tb/s Throughput</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hot-swap PSU / Fans</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1RU, 112 - 242W</td>
</tr>
<tr>
<td>GS0018</td>
<td>Top of Rack Fabric Switch</td>
<td>18 x 40Gb QSFP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replaceable Unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5Tb/s Throughput</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hot-swap PSU / Fans</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1RU, 112 - 202W</td>
</tr>
</tbody>
</table>
Switch Management

- 1GbE Out of Band Management Plane
  - PowerPC CPU - Linux
  - 1U Fixed – Controller per chassis
  - Modular - Dual redundant controllers per chassis

- Software
  - Protocol Stack
    - Pre-qualified
  - Single-point control for entire network
    - Nominated Primary with hot standbys
    - Automatic fail-over
    - Published address

- Management Interface
  - SNMP v2, v3
  - Socket interface via proxy and API
Performance

- Latency
- Scalability - Fairness
- Throughput – Bandwidth
- Application
# Performance - Latency

Gnodal hardware switching performed by switch ASIC

- **Minimum Latency** (Measured per IETF RFC1242)

<table>
<thead>
<tr>
<th>Ports</th>
<th>Hops</th>
<th>Gnodal 10GbE</th>
<th>Gnodal 40GbE</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>1</td>
<td>150ns</td>
<td>150ns</td>
</tr>
<tr>
<td>Inter switch</td>
<td>+1</td>
<td>66ns</td>
<td>66ns</td>
</tr>
<tr>
<td>720</td>
<td>3</td>
<td>282ns</td>
<td>282ns</td>
</tr>
<tr>
<td>6480</td>
<td>5</td>
<td>414ns</td>
<td>414ns</td>
</tr>
</tbody>
</table>

- **Maximum Latency** (with end-point congestion, heavily loaded)

<table>
<thead>
<tr>
<th>Ports</th>
<th>Hops</th>
<th>64 byte frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>720</td>
<td>3</td>
<td>546ns</td>
</tr>
<tr>
<td>6480</td>
<td>5</td>
<td>942ns</td>
</tr>
</tbody>
</table>

* 0.25m cable

Shared with HPCAC Lugano – Gnodal Ltd 2012
## RDMA Latency

<table>
<thead>
<tr>
<th>Bytes</th>
<th>Latency(us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>0.92</td>
</tr>
<tr>
<td>8</td>
<td>0.92</td>
</tr>
<tr>
<td>16</td>
<td>0.93</td>
</tr>
<tr>
<td>32</td>
<td>0.94</td>
</tr>
<tr>
<td>64</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Multi-Hop Latency

Latency Difference Between Nearest and Next Nearest Neighbour on a 720 Port Network

Average latency difference of 146ns
Distributed Fairness

- Achieved Bandwidth between TCP Tx and Rx in the presence of congestion
  - 512 port fat tree
  - 24 switches (16 leaf, 8 spine)

- Traffic pattern
  - Congesting incast
    - 106 ports (chosen at random) sending to 3 ports on two (left hand) leaf switches
  - Uncontending background traffic
    - 406 ports, randomly pairing, sending number frames of varying size
    - Occasionally re-pairing (random)
Performance - Bandwidth

Performance Characteristics During End-point Congestion

Typical Ethernet Switching

- Max: 46%
- Avg: 10%

Proprietary Switching

- Max: 77%
- Avg: 28%

Gnodal Ethernet Switching

- Max: 94%
- Avg: 86%

Gnodal Ethernet Switching with Anti-congestion

- Max: 100%
- Avg: 99%
2 Chassis Test Bench (Second Fabric/40GbE Link)
Adding Another Fabric Link

TCP Bandwidth (Gbps)

Number of pairs of congesting, incast streams

Gnodal
Gnodal x2
40GE
40GE best fit
3 Chassis Test Bench
Summary

- L2+ with lowest latency and power
- Highest density in 1U
  - 10GbE and 40GbE edge access
- Multi-path capability in multi-chassis installations
  - 40Gb Load balancing fabric
  - Support for millions of MAC Addresses
  - Multi-stage Scalability to 64k edge ports
- Congestion free architecture with dynamic re-routing
  - Always In Order Frame delivery
- Effective Arbitrary Size Crossbar Performance
Mont-Blanc: European HPC platform based on embedded technology

- Power consumption already limits the performance of HPC systems
  - Energy efficiency is the key to higher performance

- Mont-Blanc aims to develop European exascale approach based on embedded power-efficient technology
  - Objective 1: Develop a prototype system based on embedded technology
    - Competitive with #1 in Green500 list
    - 2014: Scalable to 50 PFLOPS on 7 Mwatts (7 GFLOPS / Watt efficiency)
  - Objective 2: Design a next-generation system to lead the Exascale race
    - Competitive with #1 in Top500 list
    - 2017: Scalable to 200 PFLOPS on 10 Mwatts (20 GFLOPS / Watt)
  - Objective 3: Develop a suite of exascale-class applications
    - Rely on OmpSs task-based programming model for scalability and efficiency

This project and the research leading to these results has received funding from the European Community’s Seventh Framework Programme [FP7/2007-2013] under grant agreement n°288777.