Focal Point

A Low-Latency, High-Bandwidth Ethernet Switch Chip
Company Overview

Fabless Semiconductor Company (50+ people)

Formed out of Caltech (1/00)

Shipping two low-latency product families today

Backed by top-tier investors

www.fulcrummicro.com
FocalPoint: an Ethernet Switch Chip

The world’s most powerful Ethernet switch chip

- Highest port density (24 10GE ports)
- Lowest latency (200ns)
- Highest performance (240Gbps)
- Most power efficient (<150mW/Gbps)
- Most integrated (single chip)
- Most scalable (fat trees, 1,000s of ports)

FocalPoint Evaluation Platform
(The world’s most integrated 10G Ethernet system)
Agenda

Datacenter Interconnect Requirements
FocalPoint Chip
FocalPoint in Datacenter Applications
Problem: Disjointed Datacenter Inhibits Scale

Multiple interconnects create islands of specialization

- **Network technologies in today’s data center:**
  - Cluster: Optimized for low latency (Infiniband)
  - Data: Low latency, robust delivery (Fibre Channel)
  - Comms: Secure, flexible, cheap, interoperable (Ethernet)

- **Ethernet is the industry’s preferred choice**
  - Poor latency characteristics led to specialized solutions
Enabling Low-Latency Fabrics

Solutions balance additive contributors to latency

Three contributors to switch latency:

1. **Store-and-forward latency** *(last bit in to first bit out)*
   - Typical vendor: 3µS per 10GE switch hop
   - FocalPoint: 150nS per 10GE switch hop

2. **Packet serialization time**
   - Typical: 0.8nS/byte at 10GE and 8nS/byte at 1GE
   - FocalPoint cut-through: 50nS (packet independent)

3. **Scheduling latency**
   - Effects store-and-forward and cut-through equally
   - Linearly dependent on egress port load
   - Solution: add more ports
     (FocalPoint has 24, others have 20 or less)
Latency and Performance Under Load

Functional Ethernet never much more than half loaded

Latency Comparison 0% Load (Cut-through vs. Store-and-forward)

Collision Avoidance vs Egress Load
Performance Comparison

Switch latency should be 10-20% of system latency

<table>
<thead>
<tr>
<th>Comparison Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
</tr>
<tr>
<td>- 16 servers per rack switch</td>
</tr>
<tr>
<td>- 20P and 24P switches</td>
</tr>
<tr>
<td>- 4 or 8 uplinks</td>
</tr>
<tr>
<td>- 25G total uplink BW</td>
</tr>
<tr>
<td>- 3 and 5 hop networks</td>
</tr>
<tr>
<td>Per-hop collision free</td>
</tr>
<tr>
<td>- 33% 16:4 configuration</td>
</tr>
<tr>
<td>- 67% 16:8 configuration</td>
</tr>
<tr>
<td>Store-n-Forward Latency</td>
</tr>
<tr>
<td>- 3 µS – standard vendor</td>
</tr>
<tr>
<td>- 150 nS - FocalPoint</td>
</tr>
<tr>
<td>Traffic Profile</td>
</tr>
<tr>
<td>- 40% 64B</td>
</tr>
<tr>
<td>- 40% 1500B</td>
</tr>
<tr>
<td>- 20% Even (64B,1500B)</td>
</tr>
</tbody>
</table>

### 3 Hops

<table>
<thead>
<tr>
<th>Frame Size</th>
<th>FP-CT Unloaded (µS)</th>
<th>FP-SF Unloaded (µS)</th>
<th>V-SF Unloaded (µS)</th>
<th>FP-CT Loaded (µS)</th>
<th>V-SF Loaded (µS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>(µS)</td>
<td>(µS)</td>
<td>(µS)</td>
<td>(µS)</td>
<td>(µS)</td>
</tr>
<tr>
<td>64</td>
<td>0.6</td>
<td>0.6</td>
<td>9.2</td>
<td>1.5</td>
<td>15.6</td>
</tr>
<tr>
<td>512</td>
<td>0.6</td>
<td>1.7</td>
<td>10.2</td>
<td>1.5</td>
<td>16.7</td>
</tr>
<tr>
<td>1500</td>
<td>0.6</td>
<td>4.1</td>
<td>12.6</td>
<td>1.5</td>
<td>19.0</td>
</tr>
<tr>
<td>10000</td>
<td>0.6</td>
<td>24.5</td>
<td>33.0</td>
<td>1.5</td>
<td>39.4</td>
</tr>
</tbody>
</table>

### 5 Hops

<table>
<thead>
<tr>
<th>Frame Size</th>
<th>FP-CT Unloaded</th>
<th>FP-SF Unloaded</th>
<th>V-SF Unloaded</th>
<th>FP-CT Loaded</th>
<th>V-SF Loaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>1.0</td>
<td>1.0</td>
<td>15.3</td>
<td>2.6</td>
<td>26.0</td>
</tr>
<tr>
<td>512</td>
<td>1.0</td>
<td>2.8</td>
<td>17.0</td>
<td>2.6</td>
<td>27.8</td>
</tr>
<tr>
<td>1500</td>
<td>1.0</td>
<td>6.8</td>
<td>21.0</td>
<td>2.6</td>
<td>31.7</td>
</tr>
<tr>
<td>10000</td>
<td>1.0</td>
<td>40.8</td>
<td>55.0</td>
<td>2.6</td>
<td>65.7</td>
</tr>
</tbody>
</table>

FP-CT: FocalPoint in cut-through mode
FP-SF: FocalPoint in store-and-forward mode
V-SF: Vendor (typical) 10GE product in store-and-forward mode
Unloaded: 0% load – a measure of fabric fall-through latency
Loaded: 33% load for 8 uplinks, 66% load for 4 uplinks
Port Density Enables Cost Effective Scale

Clos Architecture

Three-Tier Fat Tree

Two-Tier Fat Tree
Agenda

Datacenter Interconnect Requirements

FocalPoint Chip

FocalPoint in Datacenter Applications
FocalPoint Project Goals

The only low-latency feature-rich 10GE switch

- 24 10G Ethernet ports
- 200nS fall-through latency
- 240Gbps shared memory fabric
  - Fully non-blocking fabric
  - Full-rate multicast
- Standards compliant, feature rich
  - Good QoS and congestion mgmt
  - 16K MAC addresses
  - 4K VLAN and STP tables
- Process
  - TSMC 0.13µm FSG process
  - All standard flows
  - Fully outsourced GDS to customer ship
- < 1W per port, typical

Fulcrum proprietary IP
Two key IP blocks differentiate the product.

**Key Benefits:**

- **Nexus**
  - Terabit Crossbar
  - Gigahertz performance
  - Terabit capacity
  - Nanosecond latency
  - No power penalty

- **RapidArray**
  - Packet Storage
  - 720 MHz SRAM
  - 1200 MHz interconnect
  - 76.8 GB/s throughput
  - Scalable for larger designs

- 3 nS latency (including arbitration)
- Terabit(s) per square millimeter
- Usage based power consumption
- 2x the speed of vendor cores (same size, density, yield)
- Small block optimized
FocalPoint Hardware Architecture

- RX Port Logic
  - SerDes
  - PCS
  - MAC

- TX Port Logic
  - MAC
  - PCS
  - SerDes

- Switch Element Data Path
  - RapidArray™
  - (1MB Shared Memory)

- Management
  - Frame Control
    - Frame Lookup
  - Frame Handler
  - Statistics
  - LCI

- EEPROM Interface
- CPU Interface
- JTAG Interface
- LED Interface

www.fulcrummicro.com
FocalPoint Latency Detail

Ball-to-Ball Latency is less than 200ns

Time

0ns 50ns 100ns 150ns 200ns

Packet Handler
5-Stage Pipeline at 360MHz

Store-and-Forward
64-Byte Segments

Frame Scheduler

Packet Lookup

Pointer Manager

Modified Header

RX SERDES IP
Silicon proven

Nexus Crossbar
Faster than 1000 MHz

RapidArray Memory
720MHz

TX SERDES IP

1 30ns

2 15ns

3 15ns

4 50ns

5 10ns

6 30ns

7 3ns

8 10ns

9 3ns

10 20ns
Bridge Features in the Data Center

Complete Ethernet Feature Set

- **Bridge Features**
  - 16k MAC address entries
  - All spanning tree variants
  - Learning and aging controls

- **VLANs (IEEE 802.1Q)**
  - 4k VLAN entries
  - Double tagging (Q-in-Q)
  - Port-based flood groups
  - 4k Spanning Trees (IVL)

- **QOS**
  - Per port and shared memory watermarks
  - 802.1p – 8 priorities per port
  - Pause & packet discard
  - 100 Queues
  - Transmission selection

- **Link Aggregation**

- **Security**
  - 802.1x & MAC Address Security

- **Layer 2 classification engine**
  - Drop, Mirror, change priority

- **Statistics**
  - >1,000 64 bit counters

Clustering enhancements

- Flexible link agg -- 12 port trunks
- Fat tree support -- HW learning, aging
- Stacking -- In server clustering
FocalPoint Chip Plot

Over 100 million transistors

- Ethernet Port Logic
  - Phy (SerDes)
  - PCS
  - MAC

- Nexus Crossbar
  - Terabit capacity
  - 3ns latency

- MAC Table
  - 16K addresses

- Scheduler
  - 720 MSPS (64 byte segments) event rate

- Management
  - CPU interface

- Frame Control
  - Frame handler
  - Lookup
  - Statistics

www.fulcrummicro.com
FocalPoint Status Report

FocalPoint is in production

- Tape Out
- Parts received from the foundry
- Shipped first Evaluation Platform
- First packets sent (on the first day)
- External validation of fully-provisioned switching and 200ns latency
- Validation completed
- First customer announces FocalPoint-based product
- Production ramping

Q4, '05 → Q1, '06 → Q2, '06
Recent External Validation

*Industry-leading latency and performance, as expected*

---

**Figure 1: Layer 2 Zero-loss (≤0.001) Throughput**

Across 24 ports in a Full-mesh Configuration
as Measured by Ixia ixScriptMate 5.2

<table>
<thead>
<tr>
<th>Ethernet frame size (bytes)</th>
<th>Throughput %</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>128</td>
<td>100</td>
</tr>
<tr>
<td>256</td>
<td>100</td>
</tr>
<tr>
<td>512</td>
<td>100</td>
</tr>
<tr>
<td>1,024</td>
<td>100</td>
</tr>
<tr>
<td>1,280</td>
<td>100</td>
</tr>
<tr>
<td>1,518</td>
<td>100</td>
</tr>
<tr>
<td>4,400</td>
<td>100</td>
</tr>
<tr>
<td>9,216</td>
<td>100</td>
</tr>
<tr>
<td>10,240</td>
<td>100</td>
</tr>
</tbody>
</table>

Source: The Tolly Group, March 2006

**Figure 2: Cut-through Switch Latency**

Across 24 ports in a Full-mesh Configuration
as Measured by Ixia ixScriptMate 5.2

<table>
<thead>
<tr>
<th>Ethernet frame size (bytes)</th>
<th>Latency (nanoseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>311</td>
</tr>
<tr>
<td>128</td>
<td>261</td>
</tr>
<tr>
<td>256</td>
<td>249</td>
</tr>
<tr>
<td>512</td>
<td>249</td>
</tr>
<tr>
<td>1,024</td>
<td>249</td>
</tr>
<tr>
<td>1,280</td>
<td>251</td>
</tr>
<tr>
<td>1,518</td>
<td>332</td>
</tr>
<tr>
<td>4,400</td>
<td>256</td>
</tr>
<tr>
<td>9,216</td>
<td>236</td>
</tr>
<tr>
<td>10,240</td>
<td>243</td>
</tr>
</tbody>
</table>

Source: The Tolly Group, March 2006
Agenda

Datacenter Interconnect Requirements
FocalPoint Chip Architecture
FocalPoint in Datacenter Applications
Validated End-to-End Latency

Latency comparable to specialty fabrics

Myricom

Lowest Ethernet latency – ever!
2.4µs, application-to-application (MPI)

<table>
<thead>
<tr>
<th>Switch Vendor</th>
<th>MX/Myrinet</th>
<th>MX/Ethernet</th>
<th>OpenIB/InfiniBand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myricom</td>
<td>Fulcrum</td>
<td>Mellanox</td>
<td></td>
</tr>
<tr>
<td>Ping Pong Latency</td>
<td>2.4µs</td>
<td>2.4µs</td>
<td>4.0µs</td>
</tr>
<tr>
<td>Two-way data rate</td>
<td>2,397 MB/s</td>
<td>2,162 MB/s</td>
<td>1,902 MB/s</td>
</tr>
</tbody>
</table>

NetEffect

Lowest full iWARP latency – ever!
<10µs, application-to-application (MPI)

Solarflare

Lowest 1G Ethernet latency – ever!
<10µs to the application

More headlines coming soon…
Data Center Switch (Two-Tier Fat Tree)

- **Features**
  - 288 10GE ports
  - CX-4 and XFP line cards
  - Non-blocking architecture
  - 0.6μS port-to-port latency
  - 192,000 MAC addresses (effective)
  - Single-switch software image
  - 100% multicast bandwidth
  - Rich Ethernet L2 feature set

- **Composition**
  - 24 ports per blade
  - 36 chips per chassis

- **Extremely cost effective**

- **Significant industry interest**
Hash Efficiency (288-Port Switch)

- SA-DA hash for 8k MAC addresses
- Mesh round robin
- Each pixel is a port for 12 spine chips
- +/- 5% asymmetry
- Load independent
Memory Utilization for Multiple Profiles

Maximum memory of 36 chips

- 64B, 576B, 1500B
- Random
- Random profile
  - 40% 64B
  - 40% 1,500B
  - 20% flat distribution
- Even at 95% load, no drops of 1,500B frames
Three-Tier Fat Tree Architecture

\(~1\mu s\) latency from any port to any other port

- Spine Switches
- Leaf Switches

\(\cdots 288 \cdots\)

\(\cdots 12 \cdots\)  
(8)

\(\cdots 3,456\) non-blocking 10G user ports  
(4,608 10G user ports with 2:1 over-subscription)
Available Bandwidth in Multi-Tier Fat Trees

- 3-tier system performs within 2% of 2-tier system
- Larger frames → fewer hashes
  - Exposes hash inefficiencies

Frame Size vs Max Throughput

<table>
<thead>
<tr>
<th>Frame Size (bytes)</th>
<th>3 tier</th>
<th>2 tier</th>
<th>3 tier mixed frame size</th>
<th>2 tier mixed frame size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
"Fulcrum is betting that by eliminating the latency issues with Ethernet switching, the vast ecosystem that surrounds Ethernet will drive much-needed consolidation."

Simon Stanley, Research analyst for Light Reading's Comm Chip Insider