A Low Power Monolithic IP/MPLS Packet Forwarding Engine for Service Provider Networks.


Presenter: Paramesh Gopi
Configurable ASIC Solutions

NPU / Software Based solutions

Layer 2 Processing

Layer 3/4 Processing

OC-3 155Mbps
OC-12 622Mbps
GigE 1Gbps
OC-48 2.4Gbps
OC-192/10GigE 10Gbps

1-4K Routes 4-8K Flows DiffServ
32-100K Routes 20-100K Flows DiffServ / MPLS
70-500K Routes 50-500K Flows DiffServ / MPLS

ACCESS

EDGE

EDGE

EDGE

CORE

Netw ork Pe rformance V ie w
System View
Wire-Speed Line Aggregation

OC-3c / 12c / 48c

Virtual Circuit Subscriber Cloud

PMD
SONET FRAMER

ATM SAR
GLUE

IP/MPLS Packet Forwarding ASIC(s)

Packet Memory

Route/Flow Table Memory

GTL Option

Routing Protocol (BGP/OSPF / RIP)
Table Configuration
Maintenance & Control

PoS Interface

Route Processing
Flow Processing
Admission Control
Packet Filtering
Traffic Conditioning

Packet Forwarding Path

Control Plane

Control Processor

32b 50 MHz
64b 100 MHz

64b 100 MHz
System Requirements

- **Route Processing**
  - Full Internet Feed (100K prefixes)
  - Full Multi-Layer MPLS Lookup (4 LSR prefixes)

- **Flow Classification**
  - IP Src / Dest Address, Protocol Field
  - DiffServ Code Points
  - Layer 4 TCP/UDP Port numbers
  - Layer 4 TCP/UDP Port ranges
  - 64K Flows

- **Admission Control**
  - Flow Aggregates based on Flow Classification events

- **Policy driven congestion recovery**
  - RED, WRED

- **Traffic Conditioning**
  - Fair Queueing
  - Weighted Round Robin
  - 4 Delivery Priorities / 4 drop precedences per interface

- **Maximize Connection Density**
  - CO POP space is a premium commodity

- **Minimize the power figure per rack foot**
  - Carriers are limited by their infrastructural power / heat management requirements
Packet Flow

Routing Control Plane (Protocol driven)

- Route Memory
- Flow Memory
- Packet Memory

Ingress Route / Flow Processing

- Classification and Filtering
- Admission Control
- Queue Manager Traffic Shaper
- Egress Processing

From Network / Switch Fabric

- CIDR (Longest Prefix)
- MPLS (Multi-Layer TAG)
- Destination Route / NHIP

- DSCP, MPLS (x N), IPSA/DA, TCP/UDP + Range filtering
- Flow Based, SLA Enforcement RED
- Flow Based, SLA Enforcement RED
- IP Header Checksum Context based Tagging TTL Modification < L2 FCS checking >
Micro-Architecture
Functional Partitioning

Block 1
Control / Config

- Scheduler Weights
- Queue Setup and monitoring

- CPU Interface Logic
  - Route Processor
  - Flow Processor
  - Control Config Registers

- Message Interface Logic
  - Extraction Parameters
    - Context Control
    - Layer 2 Control

Block 2
Queuing

- Configurable Scheduler

- Link Manager

- Message Interface Logic

Block 3
External interfaces

- GTL / TTL External High Speed Interface to External Tables

- 64b 100 MHz Packet Memory Interface

- 64b 100 MHz Extended Table and Cascade Interface

Block 4
Network Layer Functions

- Layer 3 / 4 Messaging Fabric

- Field Extractor

- Rx FIFO
  - Tx FIFO

- Layer 2

- Multiple Interface Layer Blocks
  - OC-3c granularity

16b 50MHz PoS PHY Interface

32b 50 MHz Memory Mapped CPU Interface
64b 100 MHz Messaging Fabric
64b 100 MHz Extended Table and Cascade Interface

Block 1
Control / Config

- Scheduler Weights
- Queue Setup and monitoring

- CPU Interface Logic
  - Route Processor
  - Flow Processor
  - Control Config Registers

- Message Interface Logic
  - Extraction Parameters
    - Context Control
    - Layer 2 Control

Block 2
Queuing

- Configurable Scheduler

- Link Manager

- Message Interface Logic

Block 3
External interfaces

- GTL / TTL External High Speed Interface to External Tables

- 64b 100 MHz Packet Memory Interface

- 64b 100 MHz Extended Table and Cascade Interface

Block 4
Network Layer Functions

- Layer 3 / 4 Messaging Fabric

- Field Extractor

- Rx FIFO
  - Tx FIFO

- Layer 2

- Multiple Interface Layer Blocks
  - OC-3c granularity

16b 50MHz PoS PHY Interface

32b 50 MHz Memory Mapped CPU Interface
Micro-Architecture
Pre-Emptive Multi-Field Range based classifier

Field Extractor (Offset and Bit Mask)

HIGH VALUE
16 - 64 bit

LOW VALUE
16 - 64b

MIN COMPARE LOGIC
MAX COMPARE LOGIC

DECISION LOGIC

VALUE DECODE

From L3 / L4 Logic via Message Fabric

32 - 128 bit Key Assembled from packet

Mode Select and Enable based on data field configuration

All Data Paths are 10ns pipelined

256 Blocks deep

Ex: Filtering based on Application port ranges

TCP / UDP Layer 4 Port \textit{Min} 16b
TCP / UDP Layer 4 Port \textit{Max} 16b

HotChips 2000
Entridia Confidential
Micro-Architecture
Extended Route / Flow Table Processing (Off Chip)

Typical External CAM Row Organization for 64 K Entries

On-Chip Block 3

Class Field / Mask Select

Key Transaction FIFO (64b X 256)

IP SA 32 b
IP DA 32 b
TCP / UDP Port 16 b
DiffServ 8 b
Protocol 8 b
MPLS Tag 32 b

Value Label 41 b

Value Label FIFO (64b X 256)

32b NHID / PVC ID
Rest of Tag is used for Queuing and scheduling

64b transaction every 10ns

External SRAM

HotChips 2000
Entridia Confidential
Packet is Queued and Link is created only after classification is complete.

Dashed paths indicate Message Fabric transfers.

Scheduler orders links based on flow label. Weights are programmed via CPU.
Packets are queued on a per flow / per class basis. Each Queue receives a share of bandwidth proportional to 

\[
\text{Data Rate} \times \frac{\text{Weight}}{\text{Total Sum of Weights}}
\]
Micro-Architecture
Floorplan

- Pre-Emptive Multi-Field Packet Analyzer
- CPU Interface (Config & Mgmt)
- Network Interface Ports (PoS)
- Queue Manager and Traffic Shaper
- Fabric Control
- Work Conserving TDM Fabric
- Pre-Emptive Multi-Field Packet Analyzer
- CPU Interface (Config & Mgmt)
- Network Interface Ports (PoS)
- Queue Manager and Traffic Shaper
- Fabric Control
- Work Conserving TDM Fabric
## Silicon Validation

### Key Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>0.25um TSMC CMOS SP5M</td>
</tr>
<tr>
<td>Gate Count</td>
<td>2.5 Million</td>
</tr>
<tr>
<td>Package</td>
<td>520 HPBGA</td>
</tr>
<tr>
<td>Power Dissipation (nominal)</td>
<td>6.5W</td>
</tr>
<tr>
<td>Memory</td>
<td>ZBT SRAM (Status / Packet)</td>
</tr>
<tr>
<td>Maximum Wire-speed Throughput (40 byte packets)</td>
<td>4.3 Million packets per second</td>
</tr>
<tr>
<td>Maximum forwarding Latency</td>
<td>7 µs</td>
</tr>
<tr>
<td>Maximum Lookup table density</td>
<td>100,000 IPv4 Routes 64,000 128b Arbitrary Flows</td>
</tr>
</tbody>
</table>
Silicon Architecture  
Key Innovations

- **Packet Classification and Filtering**
  - Multi-value denominative packet classification methodology
  - 10ns pipelined non-algorithmic circuit structure

- **Dual Stage Memory Architecture**
  - Small Packet cache to maintain sustained 40byte packet transfers indefinitely
  - Main Packet memory to store large packets

- **Scalable Work Conserving TDM Messaging Fabric**
  - 6.4 Gbps
  - Single cycle (10ns) agent access
  - Single cycle arbitration
  - Deterministic transaction resolution results in extremely small latency variance
  - 10% signalling overhead