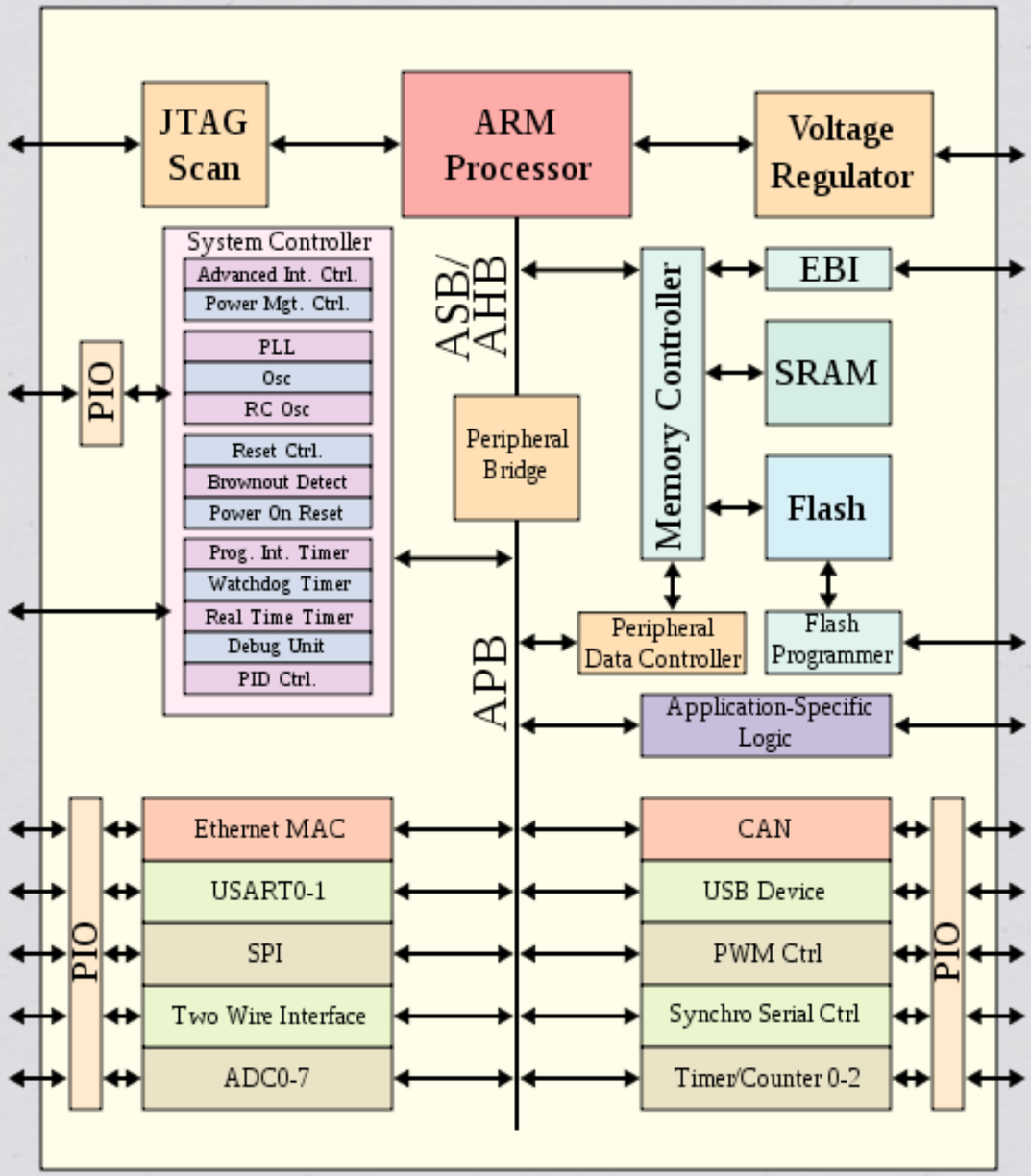
SWITCH HARDWARE

Glossary

* SOC: System-on-Chip
* CAM: Content Addressable Memory
* TCAM: Ternary Content Addressable Memory
* MII: Media Independent Interface

System-on-Chip

* Single chip with CPU, memory, flash, external interface circuitry (USB, Ethernet, Serial, etc).



Content Addressable Memory

* Memory block which can be searched for a matching data word in a single operation, returning an associative reference
* Very fast and somewhat expensive

Line	Match Data Word	Output Data
1	11000010100111110011110110011011101011010000011	000111
2	11000010100111110011110110011011000011010000100	110100
...	...	...

Line	Destination MAC	VLAN	Action
1	00:0c:29:f3:d9:ba	3715	Output: 7
2	00:0c:29:f3:d9:b0	3716	Output: 52
...	...	...	...

Ternary CAM

* Like a CAM, but supports wildcard matching - each match can now add “don’t-care” bits
* Very fast and very expensive

Line	Match Word	Output
1	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX:11010110000010000000000000000101	000111
2	10101100000010000XXXXXXXXXXXXXXXX:XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	110100
...	...	...

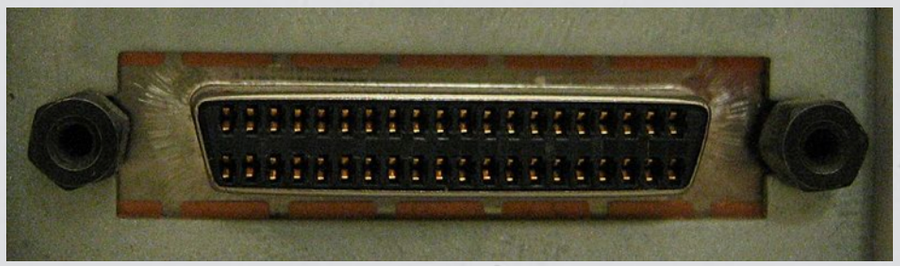
Line	Source IP	Dest IP	Action
1	*	172.16.0.5/32	Output: 7
2	172.16.0.0/16	*	Output: 52
...	...	...	...

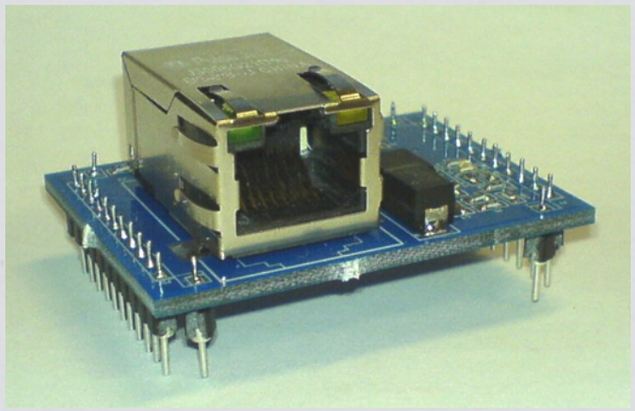
Media Independent Interface

* Standard interface used to connect devices (point-to-point) at high speeds

* Designed for FastEthernet, extended for Gigabit, 10Gb, etc.

* “Media-independent” - Copper PHY connects to exact same traces on the board as SFP, Token Ring, etc.

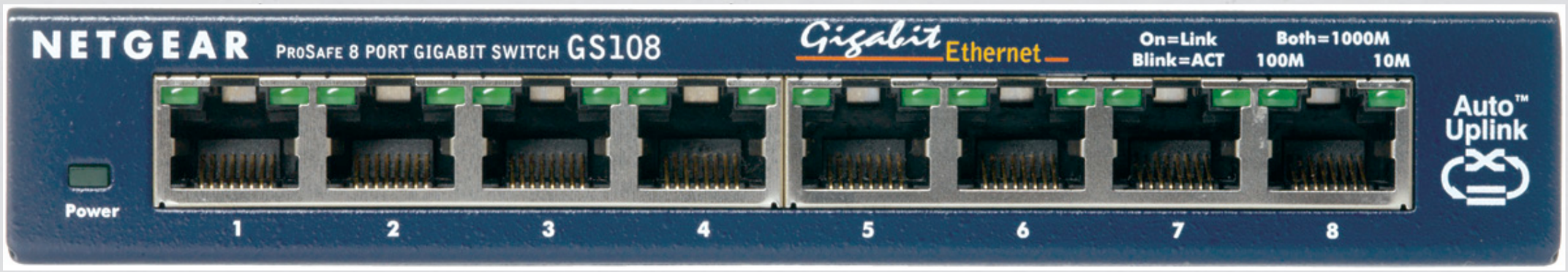




EXAMPLES

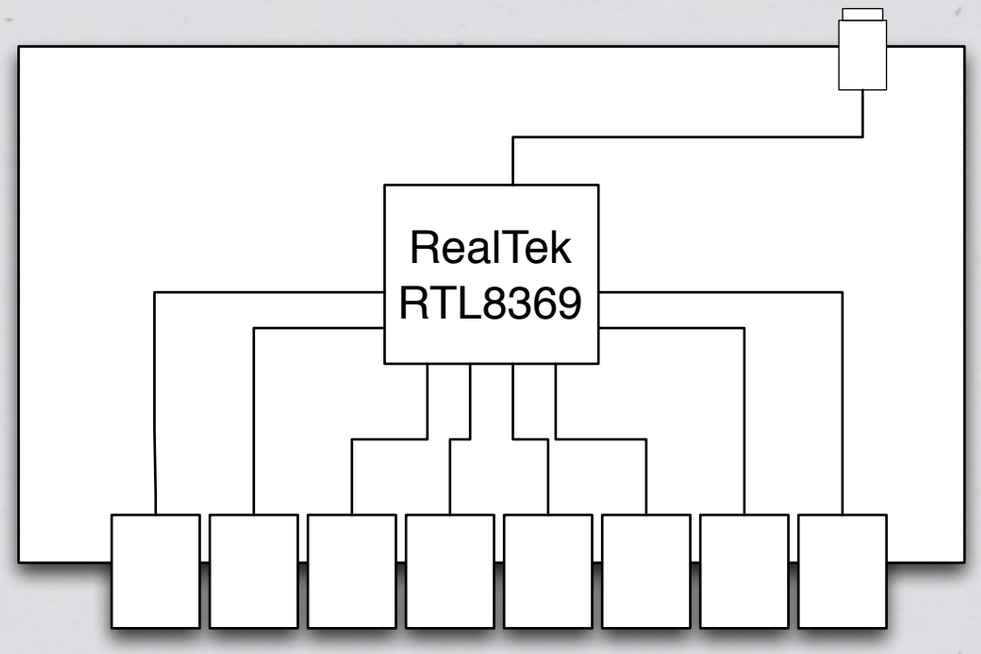
Unmanaged L2 Switch

* The most basic device currently available, a bag of ports that connects all your devices at home for cheap

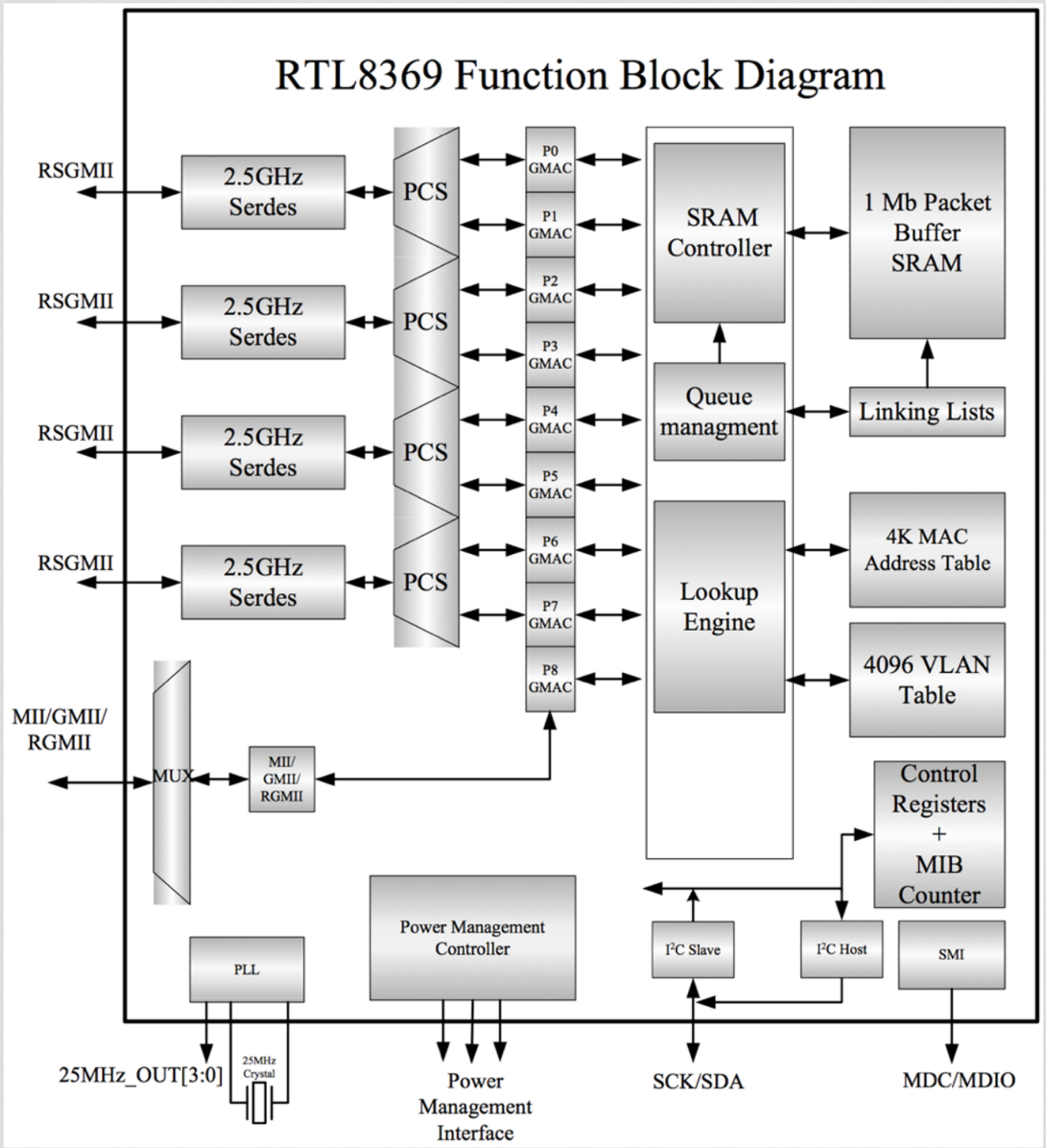


Unmanaged L2 Switch

* Single switch ASIC connected to front-panel and power
* No external management or configurability



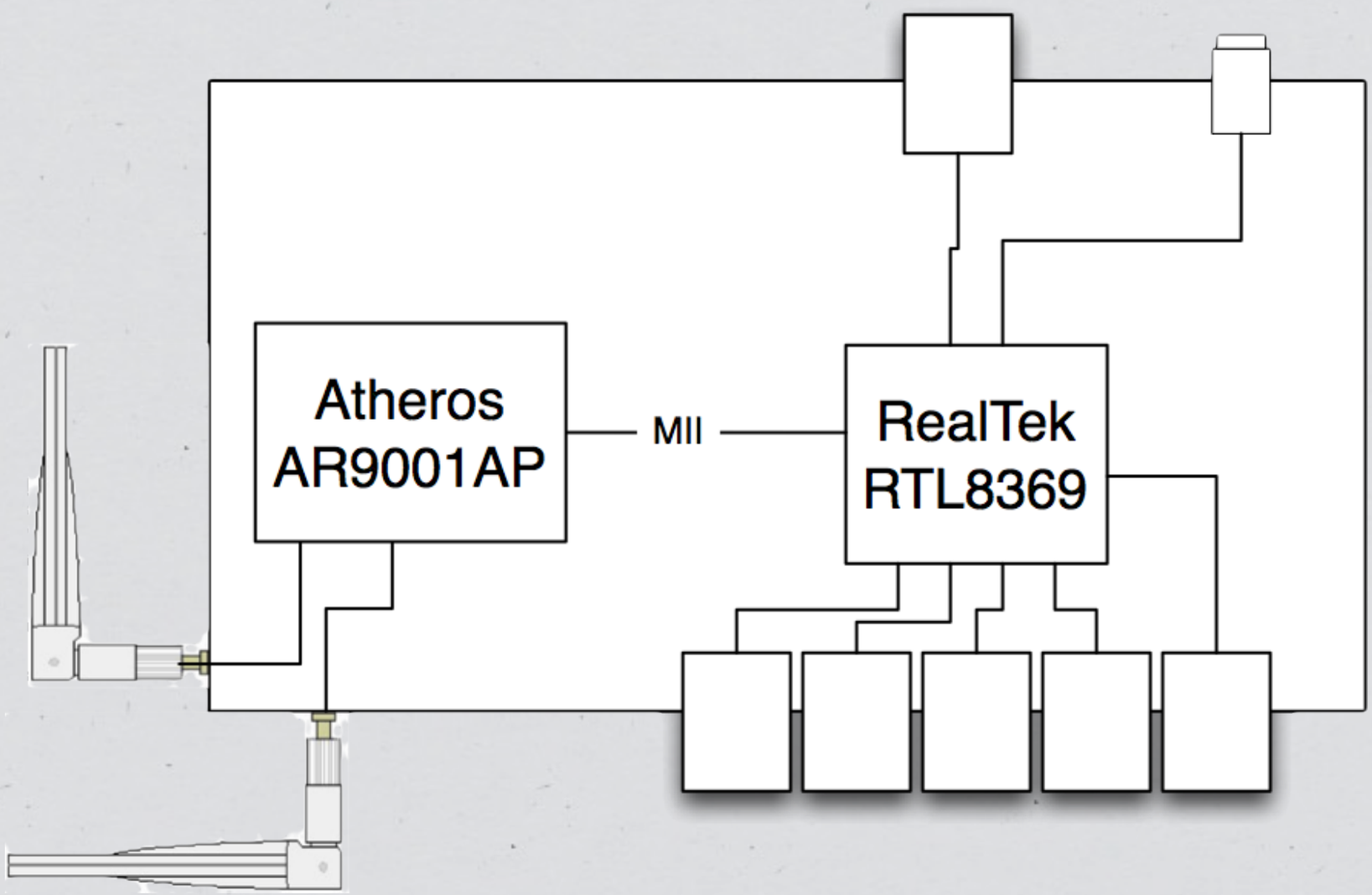
Unmanaged L2 ASIC



Wireless AP


TP-LINK
PWR
SYS
WLAN
4
3
2
1
WAN
USB
QSS
Wireless N Gigabit Router
TL-WR1043ND
QSS

Wireless AP



Managed L3 Switch

- Most common switch in use in enterprises and GENI today
  - Combination of a switch ASIC and an SOC for management
  - SOC supports management NIC, web server, SSH/Telnet to CLI, SNMP, sFlow/NetFlow, etc.
  - Configures the ASIC and polls it for statistics, but no packets ever come to the SOC
  - SOC is a small embedded processor (usually Freescale)

New/Upcoming Devices

New/Upcoming Devices

* Existing programmable merchant silicon ASICs with much bigger x86 CPUs and high-speed bus instead of embedded SoC

* Split-dataplane with wire-speed switch ASIC for most traffic, fully featured lower-throughput Network Processor for exceptional manipulations

For Something Different…

MERCHANT SILICON

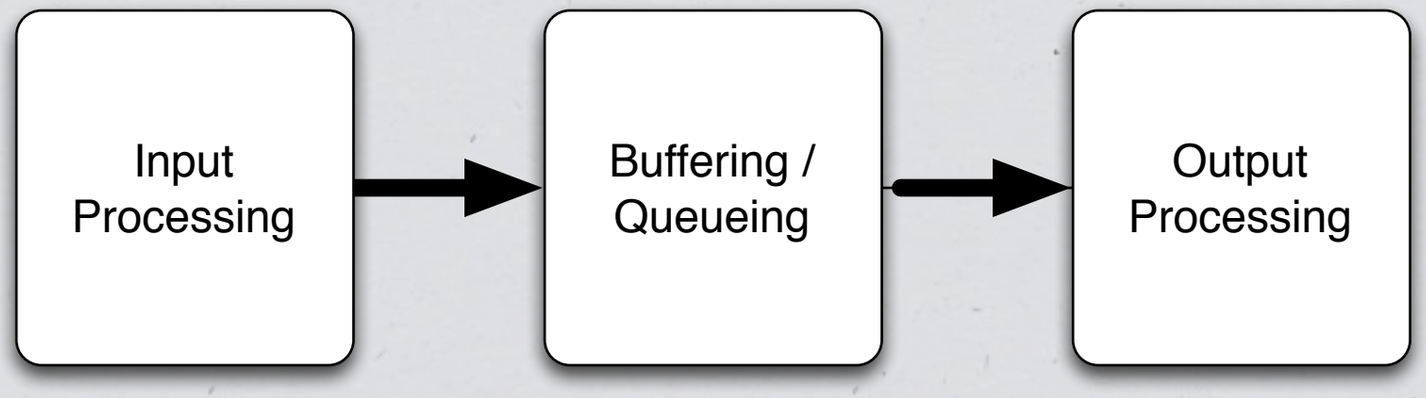
Programmable ASICs

Merchant Silicon

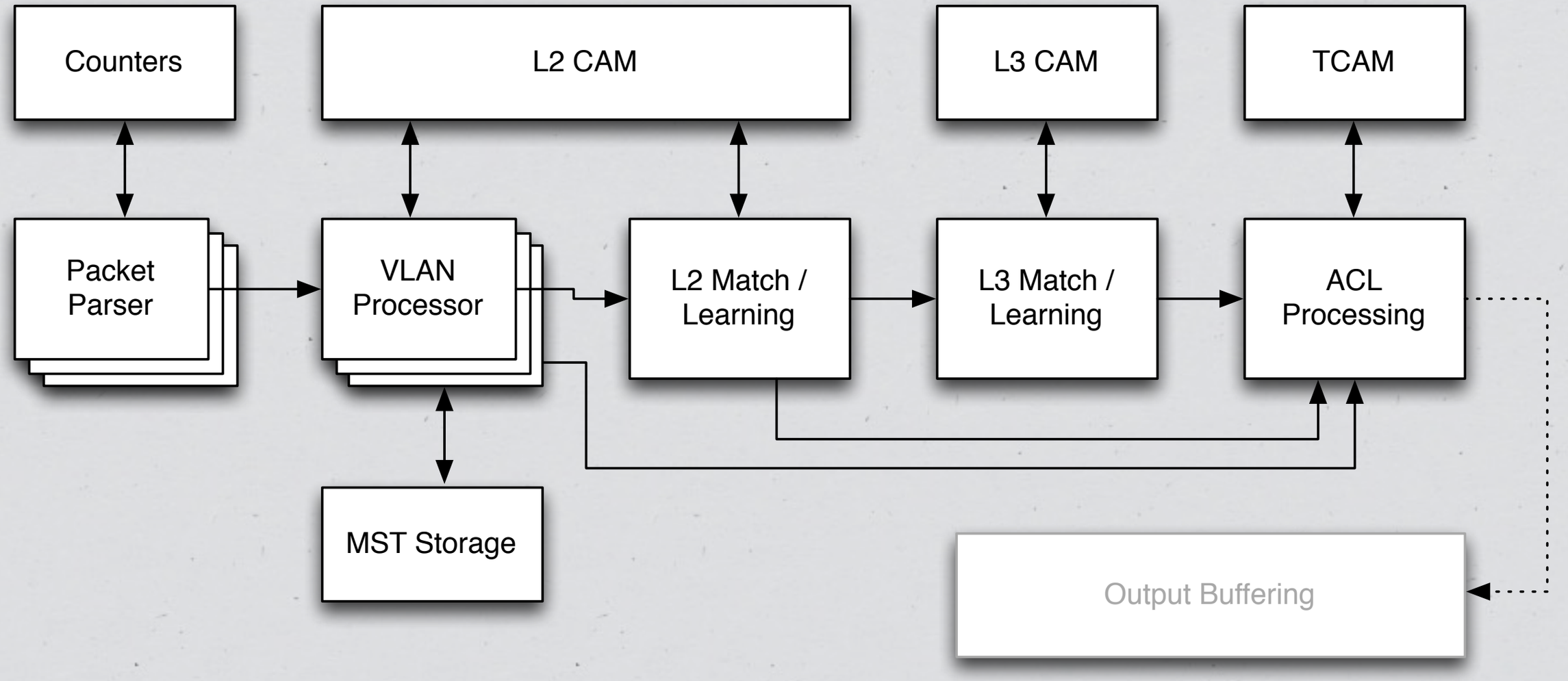
* Built by a single vendor but supplied to anyone who will buy in qty.
  * Broadcom, Marvell, Fulcrum, Centec
* Designed as general networking chips with standard throughputs and configurable feature sets
  * 176 Gbps supports 48x1Gb, 4x10Gb
  * 1.28Tbps supports 48x10Gb, 4x40Gb
* Still built with traditional networking in mind, so the flexibility only goes so far

Über-basics

General Processing Pipeline



Input Processing



Oh right, TCAMs…

* Never intended to be used for a significant portion of packet processing
  * Packets that weren't handled by the protocol support baked-in to the silicon make it here
  * A policy point for administrators to create ACLs to drop packets before output processing, or for load balancers to do L4-L7 matching before rewriting
* Making it the only available OpenFlow table is extremely limiting

Brilliant Idea!

* TCAMs seem super-useful though! Lets just make them really big and drop the rest of this pipeline nonsense!

* But wait…
  * The TCAM is the largest (by die size) and most expensive part of the ASIC
  * I suppose there's a reason no one already done this?
  * Remember how the entire TCAM needs to be searched in one operation?

TCAMs are Slow

* Fastest available TCAMs today are ~600Mhz
  * At one packet comparison per operation, that's only 600mpps
    * You need ~1.5mpps per gigabit (84 byte minimum frame size)
    * After 400Gbps things get interesting
* Now you need more TCAMs in parallel
  * Which is why TCAMs in newer devices are smaller, not larger

Output Processing

* Not as complex as input processing - mostly just needs to perform the actions that were attached during input processing
* Various ASICs support various output actions
  * Cheapest ASICs can output packets out any port, but not do any rewrite
  * Some ASICs can interleave output and rewrite actions, but most can’t

Looking Forward

- OpenFlow 1.3 adds support for performing actions mid-pipeline, but most switches can't actually do this
  - If they could, you would be introducing jitter
  - And possibly packet reordering
- IPv6 support further reduces your available TCAM space
  - TCAM field matches are somewhat configurable, so right now we are benefitting from the TCAM being configured for IPv4 src/dst