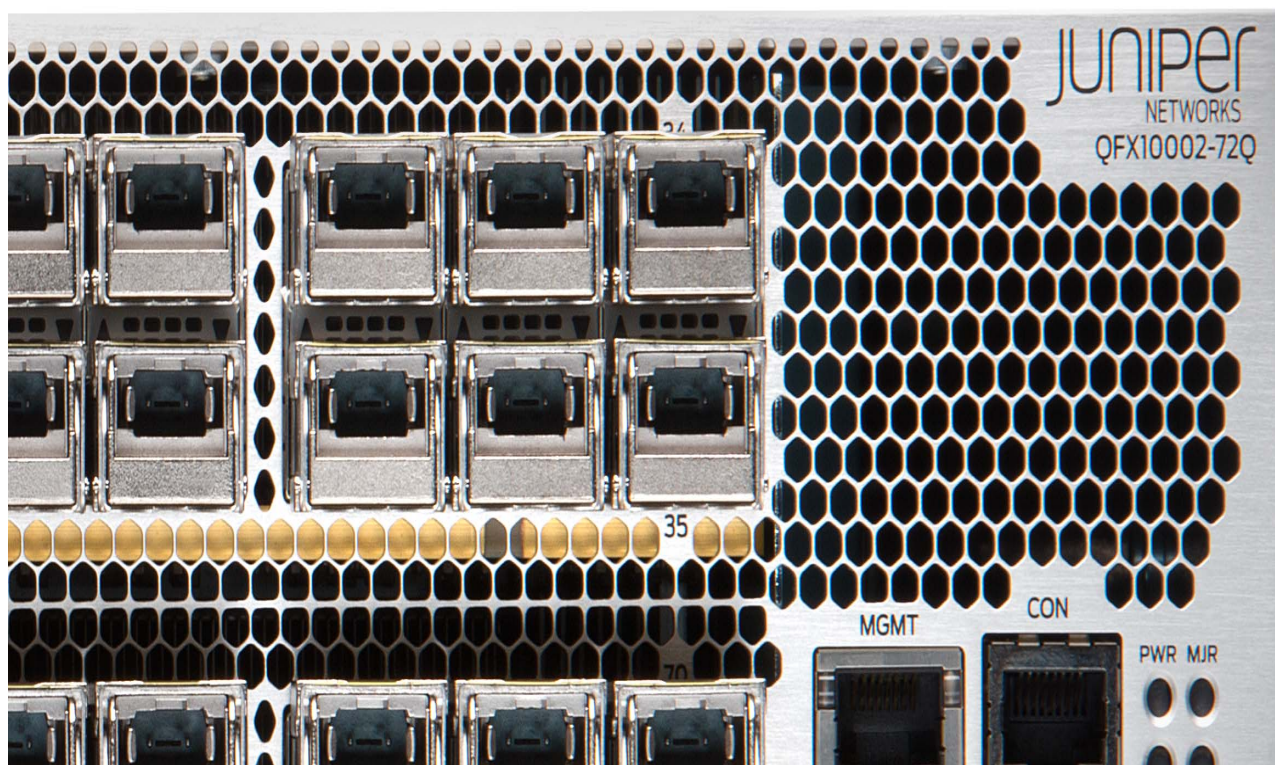




## Juniper QFX10002 Technical Overview

by  **Doug Hanks (JNPRdhanks)**  03-18-2015 09:52 PM - edited

The new Juniper QFX10002 is awesome. Let me tell you about it.



This blog post will focus on the new Juniper QFX10002. Everything mentioned in this blog is applicable to the larger Juniper QFX10008 and QFX10016. The only differences are port density between the different models of the QFX10000.

One of the first questions I'm always asked about the QFX10000 is if it's just a bunch of BRCM T2 chips. The answer is no, it's 100% Juniper silicon. The next question is if it's Juniper Trio. The answer is no, it's the new Juniper Q5 chip. For those of you who are curious, the Juniper Q5 chip got its name from being dedicated to the QFX family and each chip operates at 500Gbps full duplex. Pretty cool.

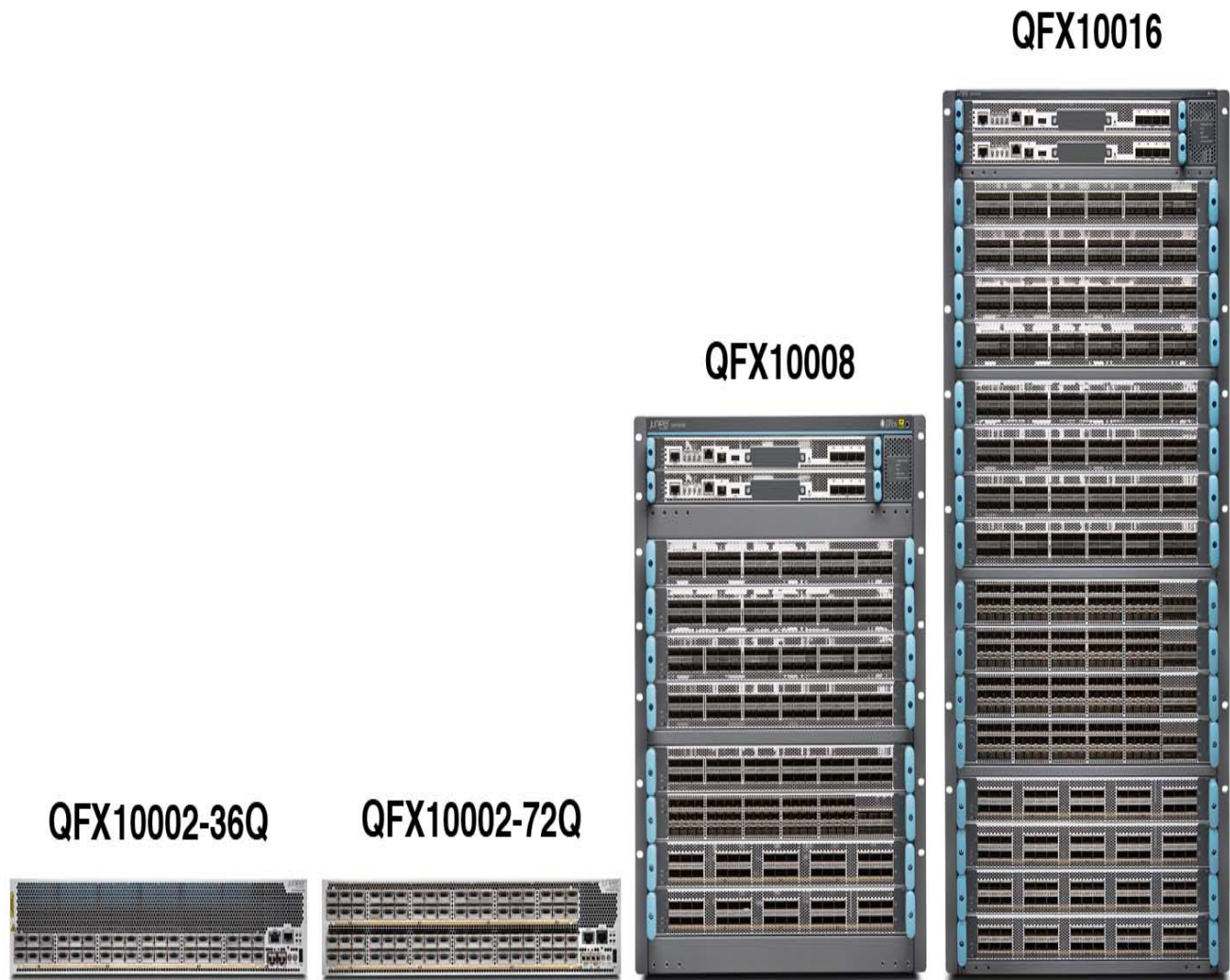
The new Juniper Q5 chip has been designed from the ground up to solve the difficult spine and aggregation challenges in the data center. Our customers have been demanding:

- High logical scale (Host, LPM, and ACLs)
- Feature rich encapsulations (VXLAN, MPLS, and GRE)
- Large buffer (to handle micro-bursts, incast, and virtual buffering)

We're not able to get all of this functionality and scale from merchant silicon, so we simply had to design the Juniper Q5 chip in-house. However we firmly believe in a strategy of combining merchant silicon and Juniper silicon in the data center. We see that merchant silicon does very well in the top-of-rack or access layer and is complimented with Juniper silicon in the core and aggregation layers of the data center.

## System Overview

The Juniper QFX10000 is a new family of switches that are designed for the core and aggregation in the data center. There are four models:



The first two switches are fixed format, and the last two switches are modular chassis.

For now let's just focus on the Juniper QFX10002. It's 2RU in size with 36xQSFP28 or 72xQSFP28 ports in the front. These can be used as either 40GbE or 100GbE. The Juniper QFX10002 will be right at home in the spine or aggregation of the data center. Obviously it can support either 36 or 72 top-of-rack switches, and on the assumption of 48x10GbE per access switch, that brings your total data center capacity to 3,456x10GbE with just a pair of QFX10002 spine switches. Not too bad for a little and mighty 2RU switch.

## Features

The Juniper QFX10000 is very feature rich. Here's a quick snapshot of what it can do:

- Network Architectures and Technology
  - IP Fabric
  - MC-LAG
  - MPLS fabric
  - Junos Fusion
- High Availability
  - Non-stop routing (NSR)
  - Non-stop bridging (NSB)
  - Graceful routing engine switchover (GRES)

- In-service software upgrade (ISSU)
- Automation Capabilities
  - Apache Thrift APIs for control plane and data plane
  - Puppet
  - Chef
  - Yocoto Linux foundation with Linux containers and virtual machines
  - Python
  - Execute any unsigned binaries on Linux or Junos
  - Ansible
  - NETCONF
- Analytics
  - Precision Time Protocol (PTP) time-stamping
  - Stream interface and bandwidth statistics through GPB, JSON, CSV, TSV
  - Application visibility and workload placement
  - Elephant flow detection and mitigation
- IP
  - Full L2 – 16K bridge domains
  - Full L3 – 256K LPM table
  - IPv4 and IPv6 support
  - Full multicast support
- MPLS
  - LDP and RSVP
  - L3VPNs
  - L2VPNs
  - EVPN
  - MPLSoUDP
  - MPLSoGRE
  - MPLSoMPLS (vanilla transport + service labels)
- VXLAN
  - L2 gateway
  - L3 gateway (routing)
  - OVSDB and EVPN control plane

## Logical Scale

The QFX10000 packs a big punch when it comes to logical scale. High-speed hybrid memory cubes give the Juniper Q5 plenty of storage space for firewall filters, MAC addresses, host entries, and longest prefix matches. Here's the Juniper QFX10002 logical scale at a glance:

- 1,000,000 MAC addresses
- 2,000,000 host routes
- 16,000 bridge domains
- 256,000 IPv4 and IPv6 FIB (yes! this is multi-dimensional scale)
- 48,000 firewall filters
- 384,000 firewall terms
- 48,000 policers
- 2,304,000 virtual output queues
- 4,096 GRE tunnels
- 4,096 MPLS VPNs

## Juniper Q5 Chip

All of the QFX10000 horsepower comes from a new chip called the Juniper Q5. We designed it in-house to specifically meet the high logical scale, big buffering, and feature rich demands coming from our customers in the data center.

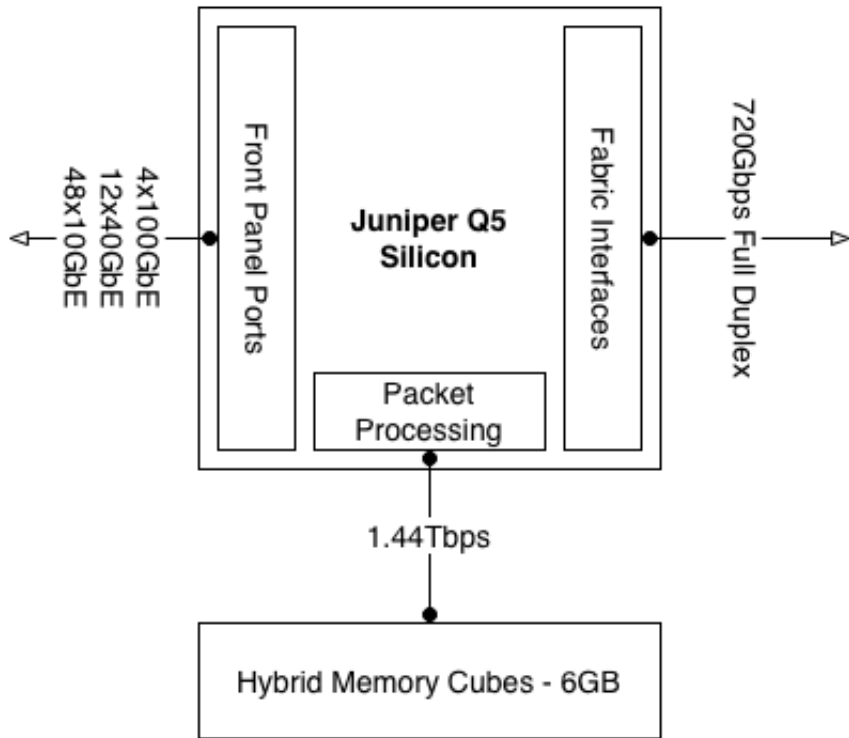


Each Juniper Q5 chip is able to process 500Gbps of full duplex bandwidth and can process 333Mpps. There are three major components to the Juniper Q5 chip:

1. Front panel ports
2. Packet processing
3. Fabric interfaces

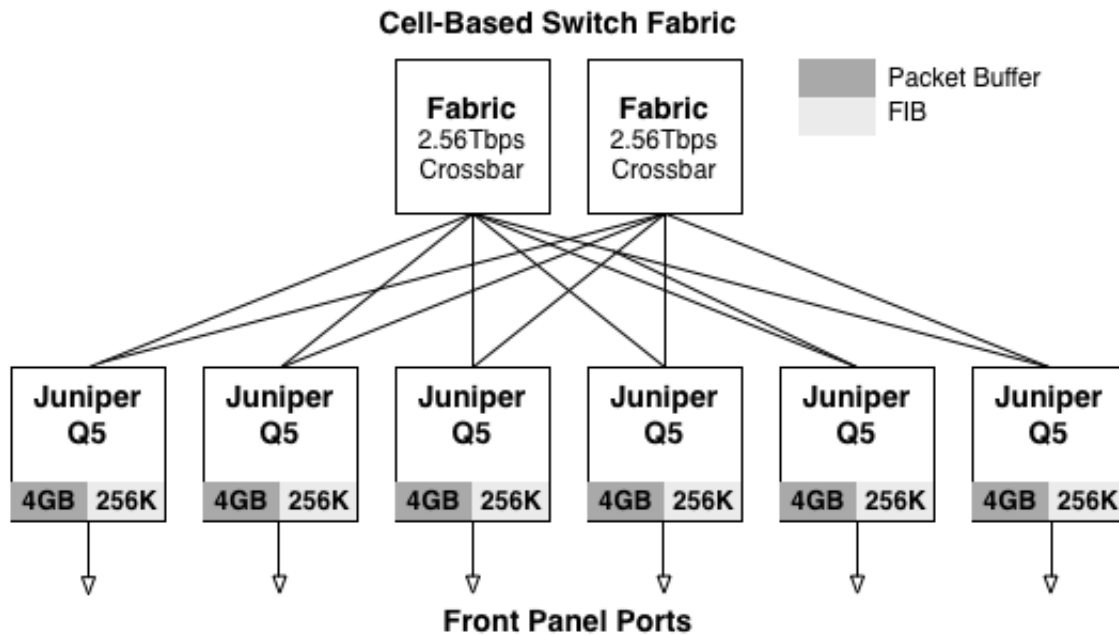
Obviously the front panel ports drive the external interfaces on the front of the switch. You can see that each Juniper Q5 chip can handle 12x40GbE interfaces, so the Juniper QFX10002-72Q has a total of six Juniper Q5 chips just to handle the front panel ports.

Packet processing has 6GB of storage available using the high-speed hybrid memory cube (HMC). 4GB is reserved for packet buffering and the other 2GB is reserved for logical tables.



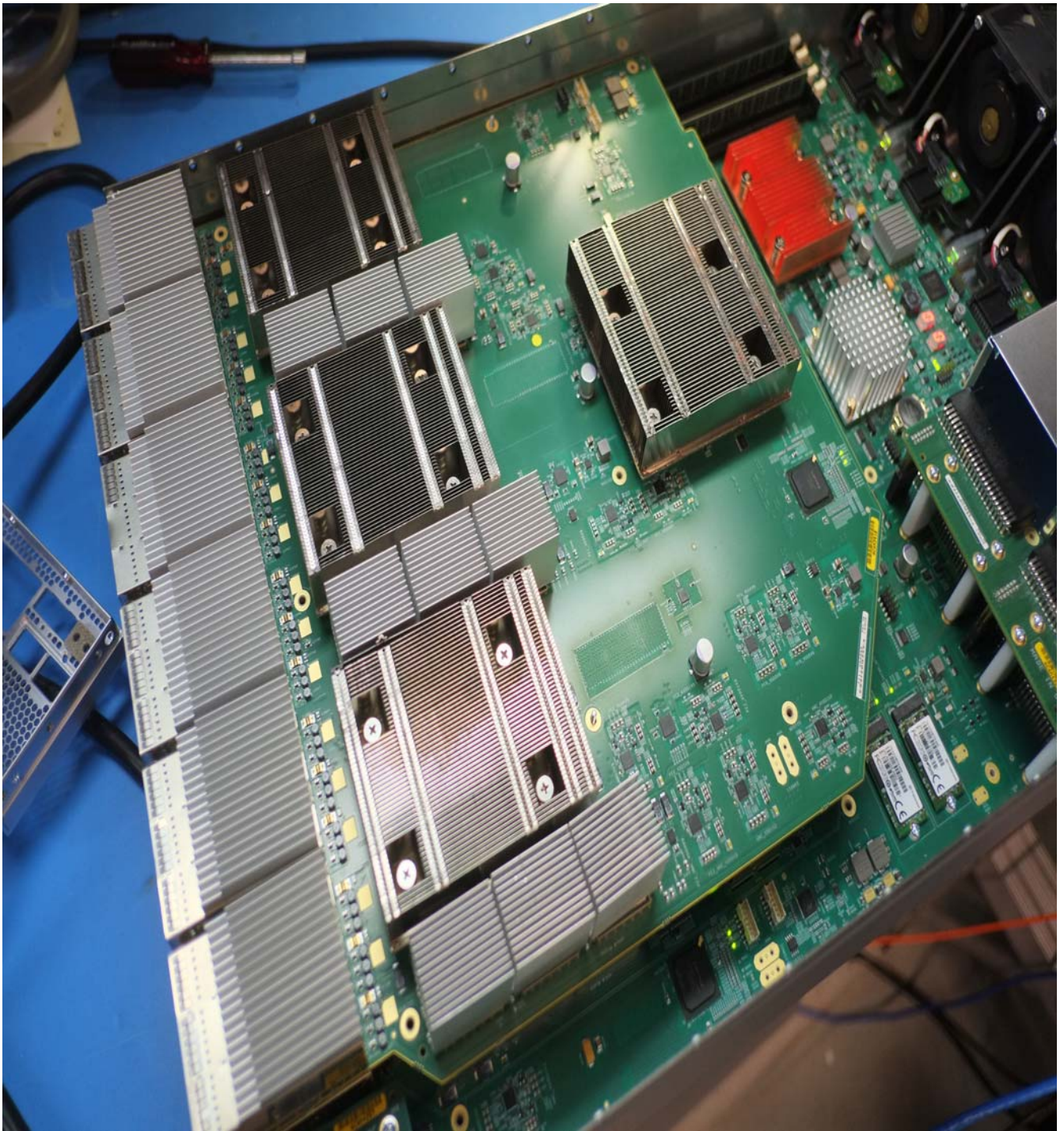
### Switch Fabric

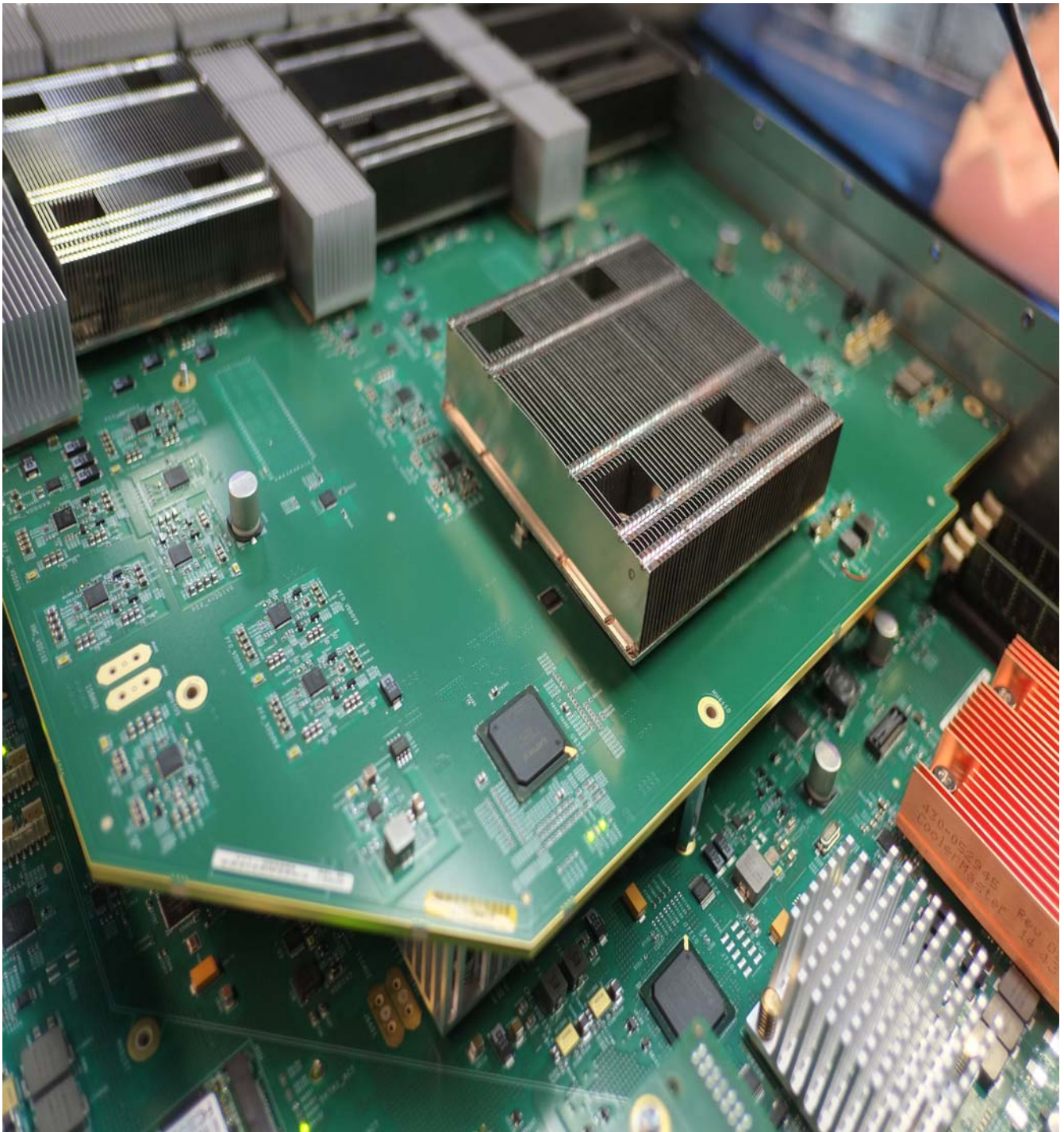
Obviously with having either three (QFX10002-36Q) or six (QFX10002-72Q) Juniper Q5 chips for front panel ports, the Juniper QFX10002 is a multi-chip system. Each Juniper Q5 chip has fabric interfaces that allow them to be combined in a cell-based switching fabric as shown below.



Two Juniper switch fabric chips are used for the cell-based switch fabric, which connects the other six Juniper Q5 chips used for the front panel ports. Each switch fabric chip can be a 192x192 crossbar that's each able to handle 2.56Tbps full duplex. The result is a non-blocking switch that's able to forward 5.76Tbps at line-rate with high logical scale, large buffers, and advanced features.

The Juniper QFX10002 is split into two mezzanines. Each mezzanine has three Juniper Q5 chips and a single Juniper switch fabric chip as shown below.





Each front panel Juniper Q5 chip breaks the front panel port traffic into cells that are sprayed evenly across the fabric. Depending on the current chip's utilization and overall bandwidth, the cell size is changed dynamically to maximize the transport efficiency across the fabric. There possible cell sizes are 96, 112, 128, 144, 160, or 176 bytes.

### Hybrid Memory Cube and Bloom Filters

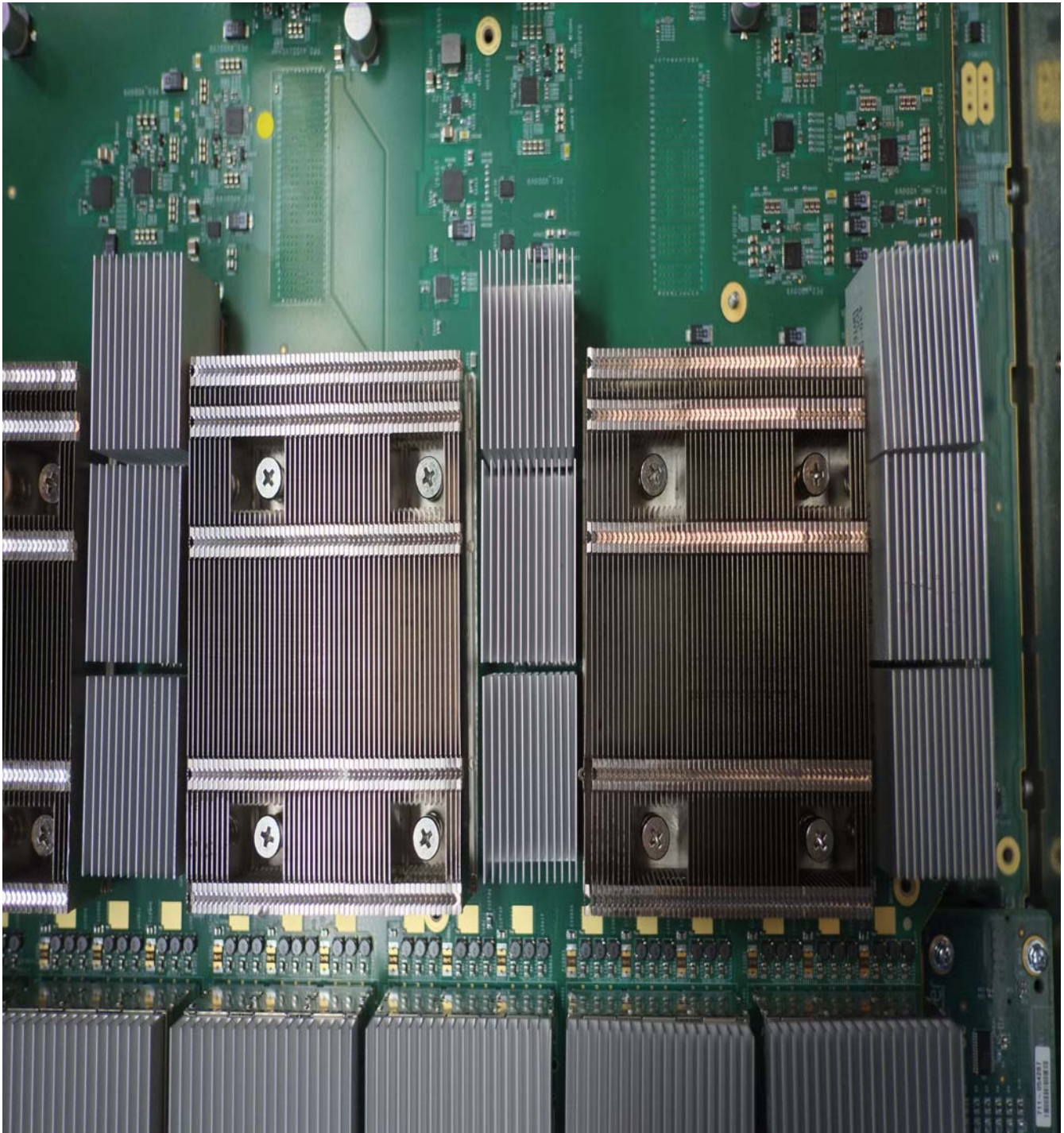
Traditionally TCAMs have been used in networking equipment to provide fast searches for tables like LPM, ACL, and others. However there are some drawbacks to TCAM: it's expensive, takes up a lot of room, and uses a lot of power. If you want to provide massive logical scale, unfortunately a TCAM isn't sufficient.

To solve the challenges of high logical scale, we need to think about two things: a new type of storage and what type of search algorithm to use. One storage option is to use DDR4 memory; it's cheap and plentiful. However DDR4 requires a lot of chips, pins and surface area when you need massive scale.

Juniper decided to use a new class of 3D memory technology called hybrid memory cube (HMC). It's purpose built

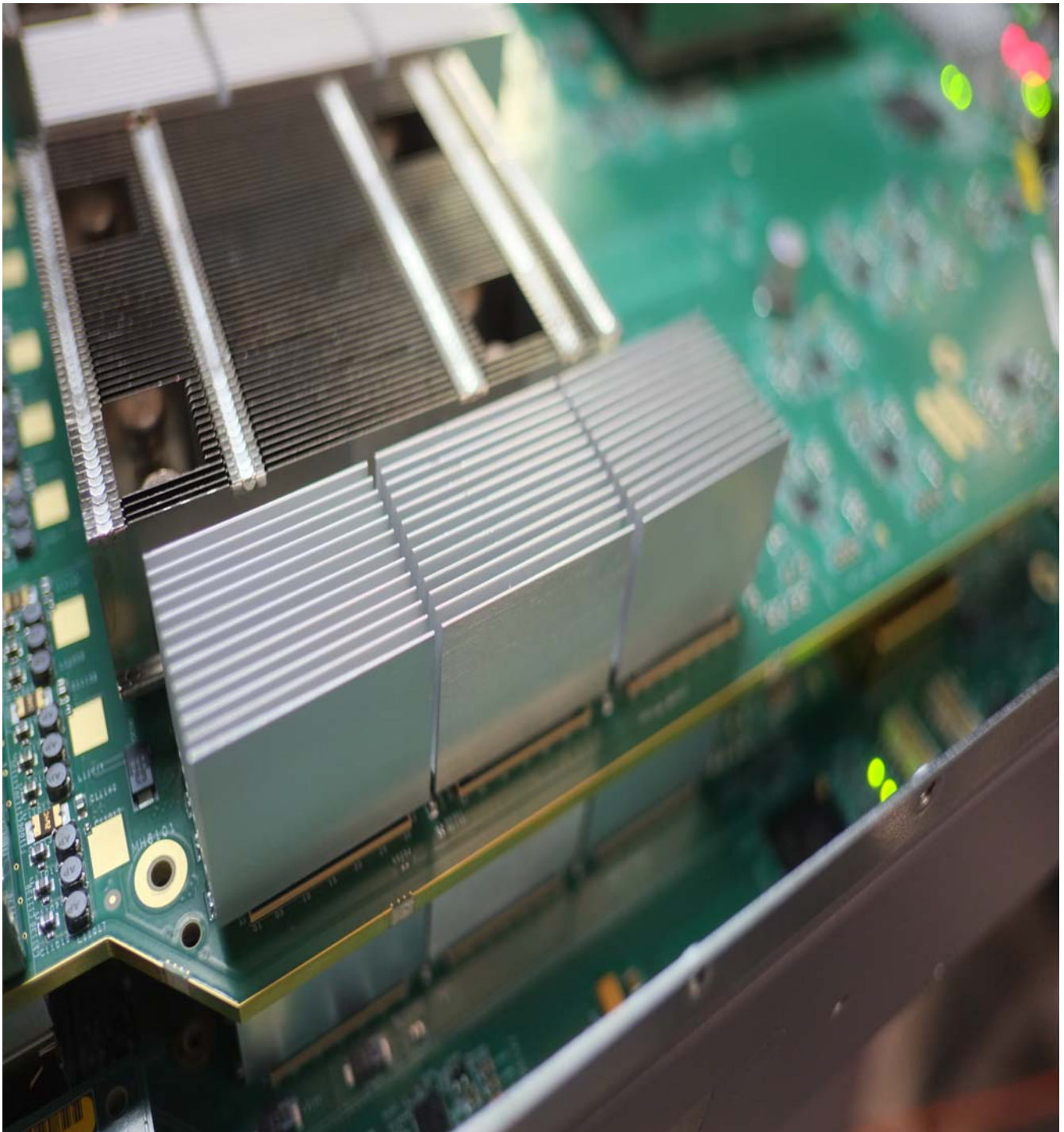
for speed, small size, and performance. To learn more about HMC, check out Salman's blog over at <http://forums.juniper.net/t5/Data-Center-Technologists/QFX10000-a-no-compromise-switching-system-wit...>

Each Juniper Q5 chip is paired with three Juniper HMC chips for storage as shown below.



Here's a closer picturing focusing on a single chip.





The next step is to think about what search algorithm to use in replace of the TCAM with our new Juniper HMC memory. We want a very efficient search algorithm that's able to result a set of results within a single lookup. We would also like to use the same search function across the entire Juniper Q5 chip for things like LPM, ACLs, and other tables.

What's surprising is that there's no TCAM in the QFX10000. The challenge using a TCAM to get the scale listed above would be extremely expensive, consume a lot of power, and take up a lot of space in the switch.

Given these limitations, we still wanted 1 read (search) per packet for LPM. Juniper's Bloom Filter Technology allows us to get 1 search for LPM in nearly all cases with a high number of entries, which enables the high logical scale.

Bloom Filter Technology works by first creating a very large bit array to store a bunch of 1s and 0s; let's call this *bloom\_filter[n]*. The next step is that when a key is inserted into the bloom filter, it is run through several different hash functions. For example if the key "192.168.1.1" was run through three hash functions and the outputs were 27,

56, and 18. We would set the corresponding offset in *bloom\_filter* equal to 1. For example the following would be set to 1:

- *bloom\_filter*[27] = 1
- *bloom\_filter*[56] = 1
- *bloom\_filter*[18] = 1

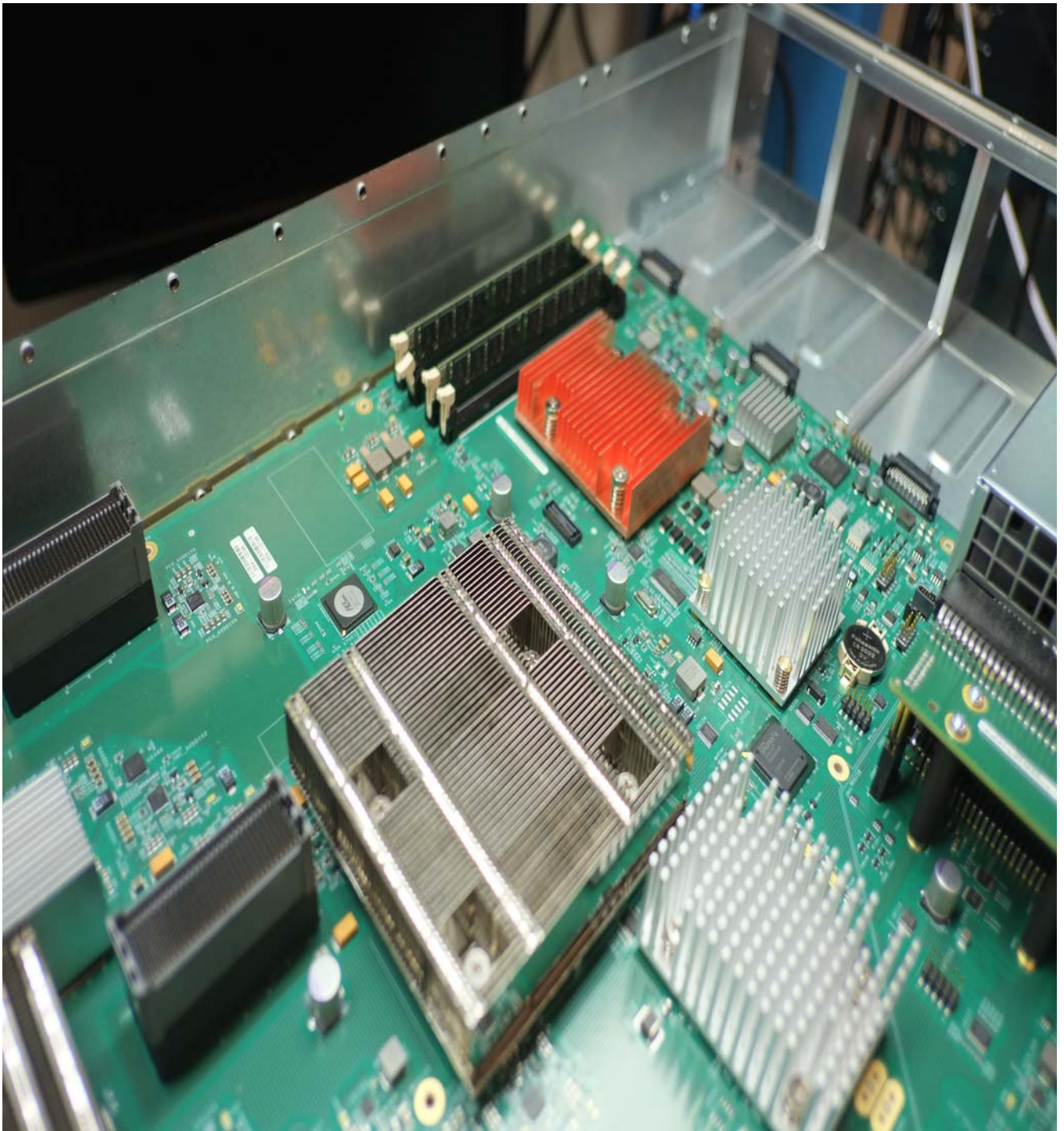
When we want to search, we simply run the search argument through the same three hash functions and compare the output to the corresponding offset in the *bloom\_filter*. If all three offsets are equal to 1, there is a high probability we have a match. However, if any of the offsets are equal to 0, we know there is definitely not a match.

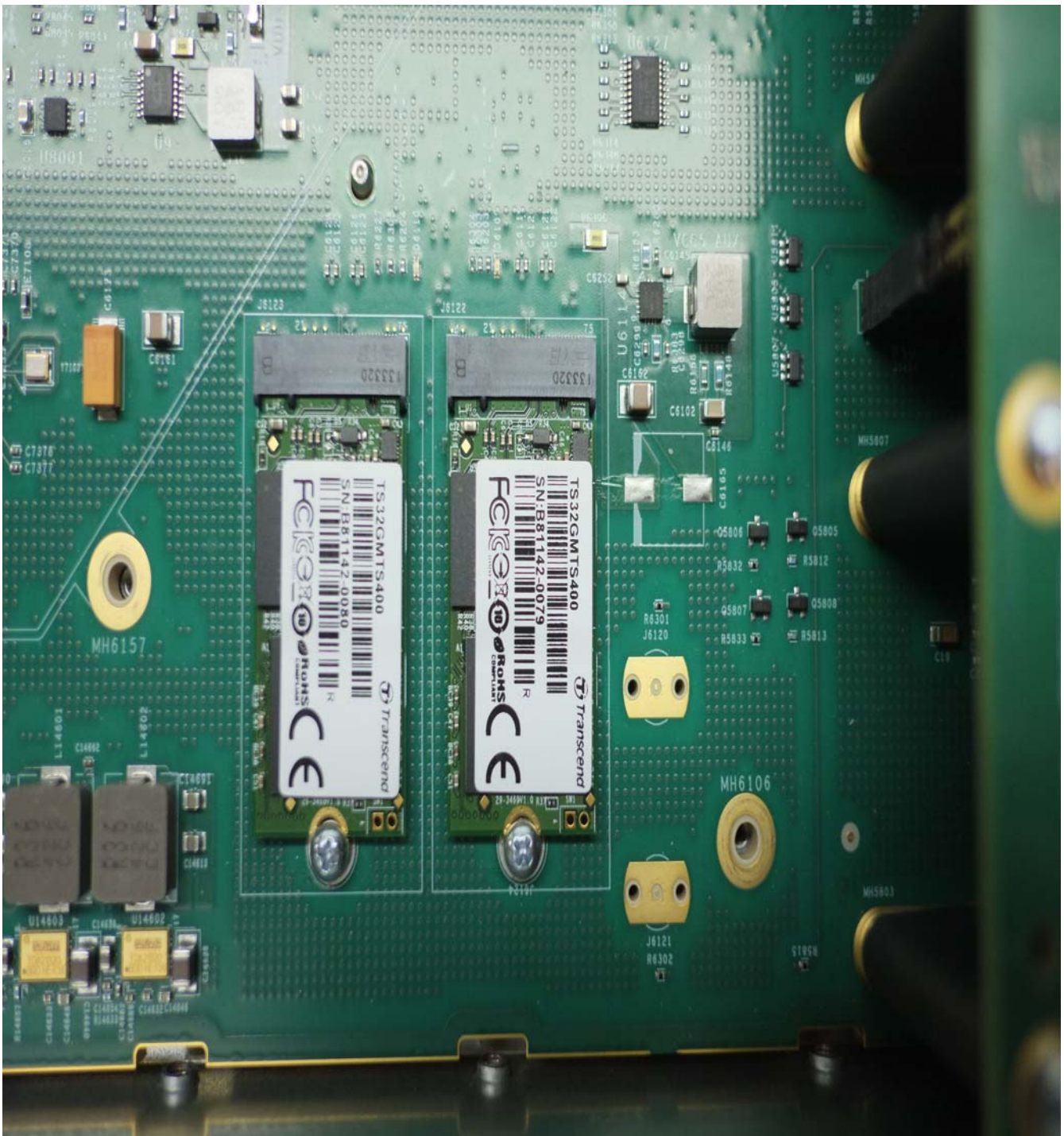
The Bloom Filter Technology uses hybrid memory cubes as the off-chip storage. There is 2GB of storage available for switching tables and bloom filters. By using a new search algorithm and high-speed memory, Juniper is able to provide high logical scale without compromising performance.

When Juniper combined a bloom filter with the new 3D HMC storage, the result is radically different and innovative way to provide massive logical scale in a data center switch.

## Control Board

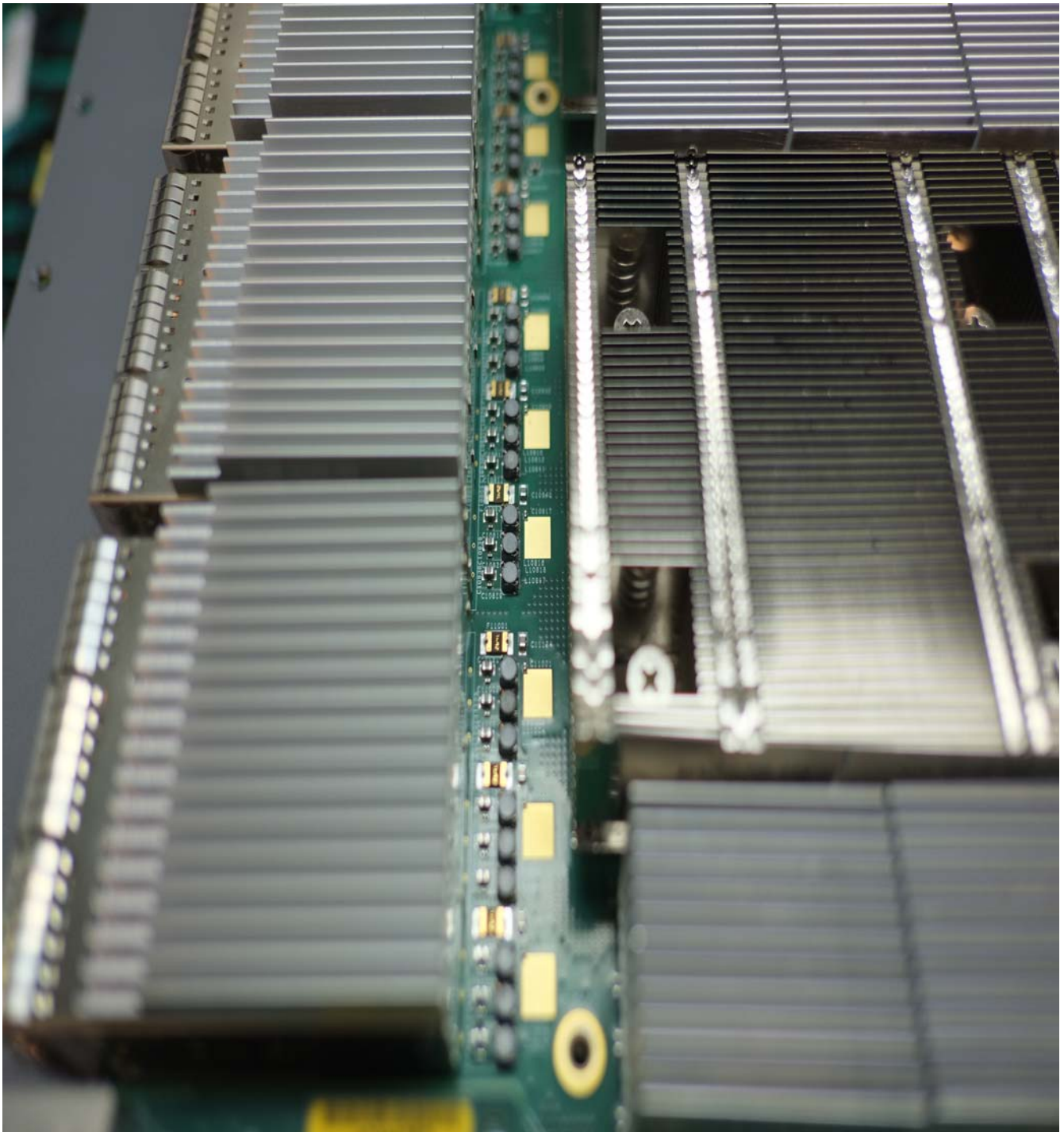
The Juniper QFX10002 is built like a server. It uses the latest Xeon processors and DDR memory. It's even using the latest M.2 SSD for storage.





## Interfaces

The Juniper QFX10002 uses a PHY-less design. The result is that the switch uses less power and provides lower latency. The front panel ports are shown on the left and the Juniper Q5 chip is shown on the right. You can notice the lack of PHYs in the middle.



The Juniper QFX10002 supports 10GbE, 40GbE, and 100GbE with standard DAC and optical interfaces without any external PHY.

Recall that each Juniper Q5 chip is able to support 10GbE, 40GbE, and 100GbE. Each interface on the Juniper QFX10002 can be configured to support any of these speeds. Each interface uses the QSFP28 form factor. The Juniper QFX10002-72Q can support the following number of interfaces:

- 288x10GbE
- 72x40GbE
- 24x100GbE

The Juniper QFX10002-36Q can support the following number of interfaces:

- 144x10GbE
- 36x40GbE

- 12x100GbE

You can simply use a breakout cable on the 40GbE interfaces to get 4x10GbE per port. You can also enable 100GbE interfaces in groups of three interfaces as shown below. For example if you enable 100GbE on interface 1, ports 0 and 2 would be disabled. If you enabled interface 5 for 100GbE, ports 3 and 4 would be disabled. You can see that ports are grouped sequentially in sets of three.

0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
36	38	40	42	44	46	48	50	52	54	56	58	60	62	64	66	68	70
37	39	41	43	45	47	49	51	53	55	57	59	61	63	65	67	69	71

■ = disabled

There's no requirement to all 40GbE or all 100GbE interfaces. You can enable 100GbE on any of the following interfaces: 1, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, 35, 37, 41, 43, 47, 49, 53, 59, 61, 65, 67, or 71.

The interface names and 100GbE interfaces are exactly the same for the Juniper QFX10002-36Q, except it excludes interfaces 36 to 71.

## Power

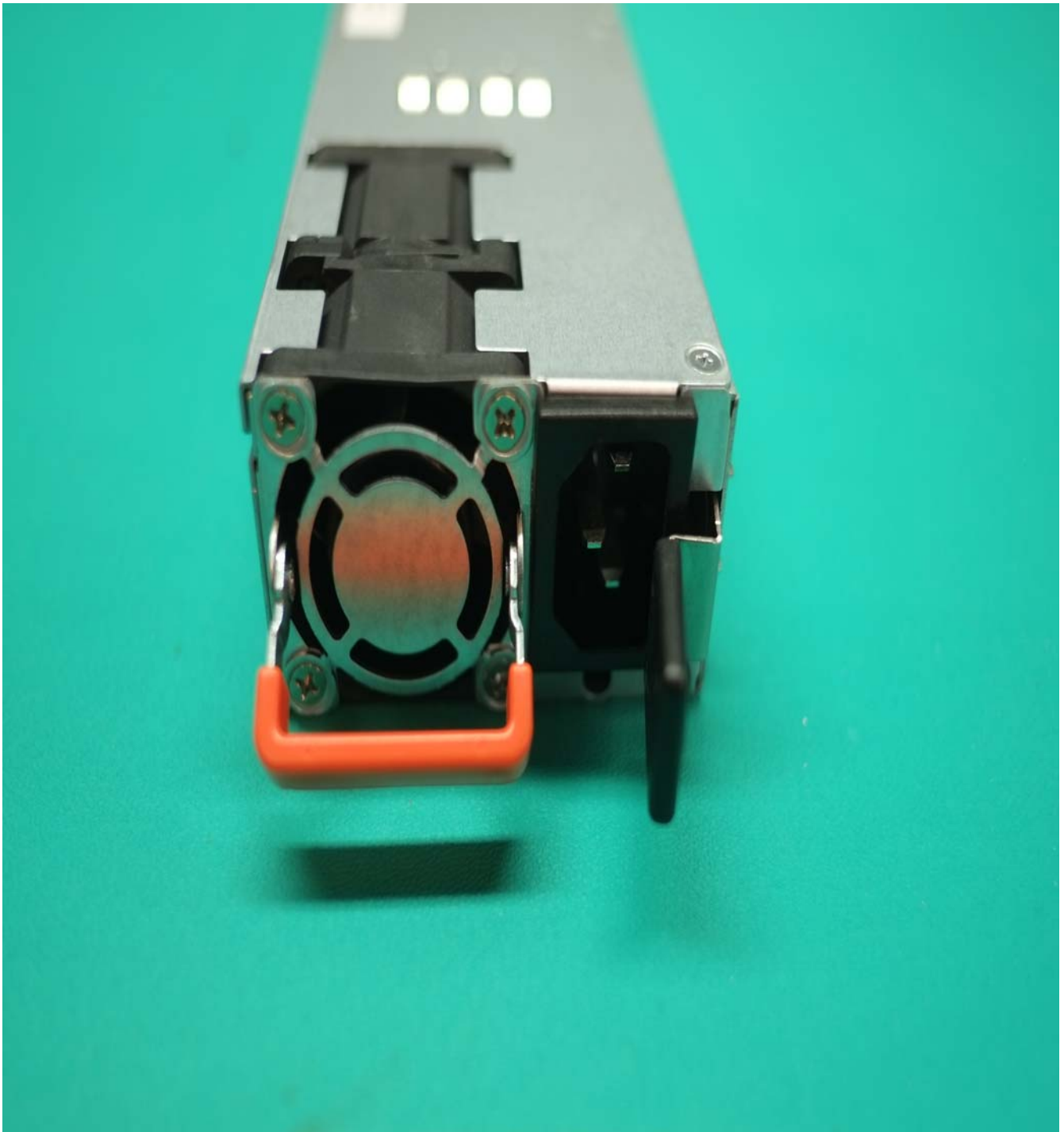
The Juniper QFX10002 is built for massive scale and is very power efficient. The typical power usage is 1300W, which is 4.5W per 10GbE port. However keep in mind the Juniper QFX10002 is built for scale, so let's put this into context. A traditional top of rack switch can support about 1,000 firewall entries and consumes about 150W of power, so each ACL consumes 150mW. The Juniper QFX10002 only uses 3mW per firewall entry, which is 98% more efficient.

The Juniper QFX10002 has four power supplies, which provide 2 + 2 redundancy.







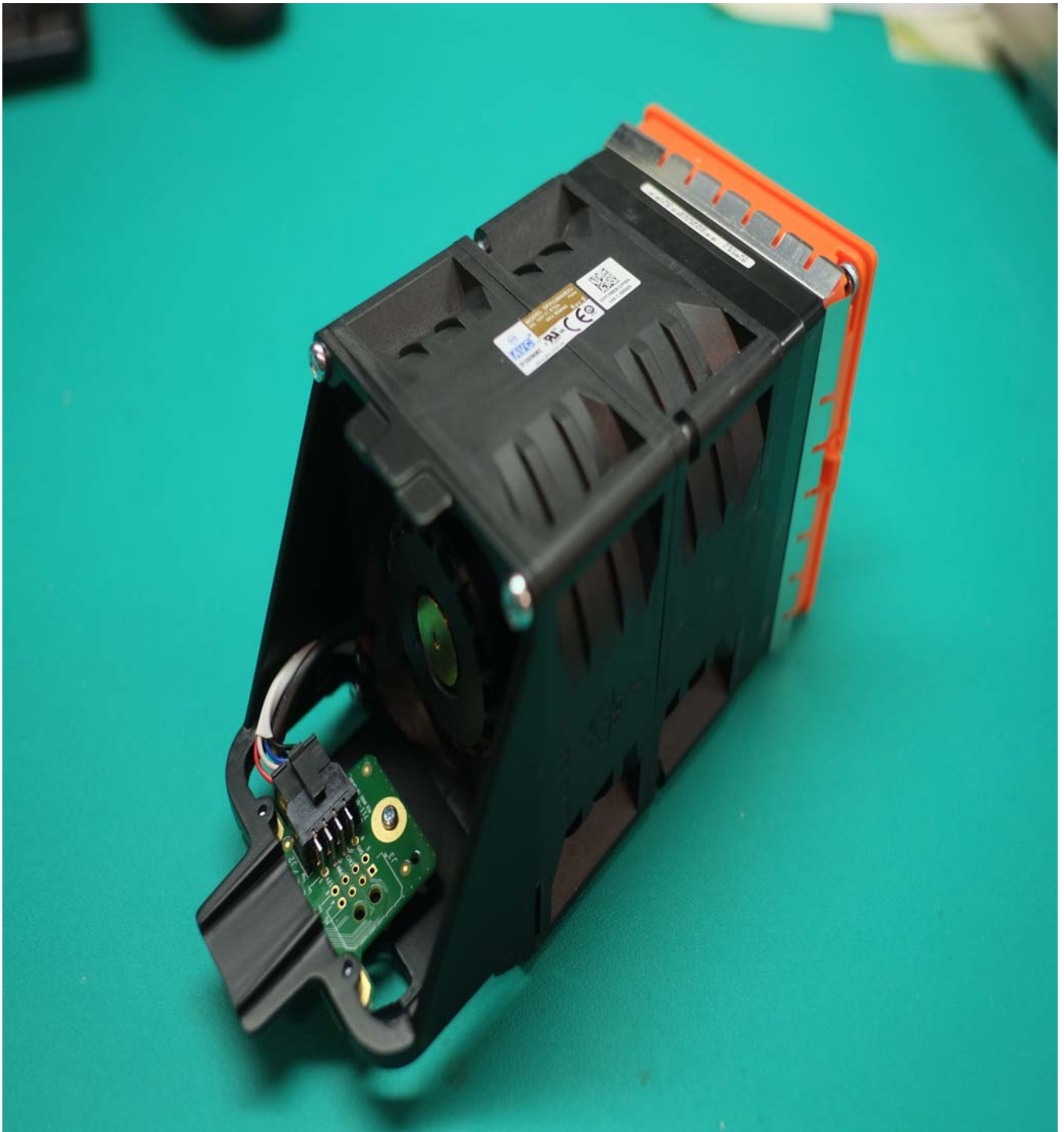


The same Juniper color scheme of orange and blue is used on the Juniper QFX10002. The orange is for front-to-back cooling and the blue is back-to-front.

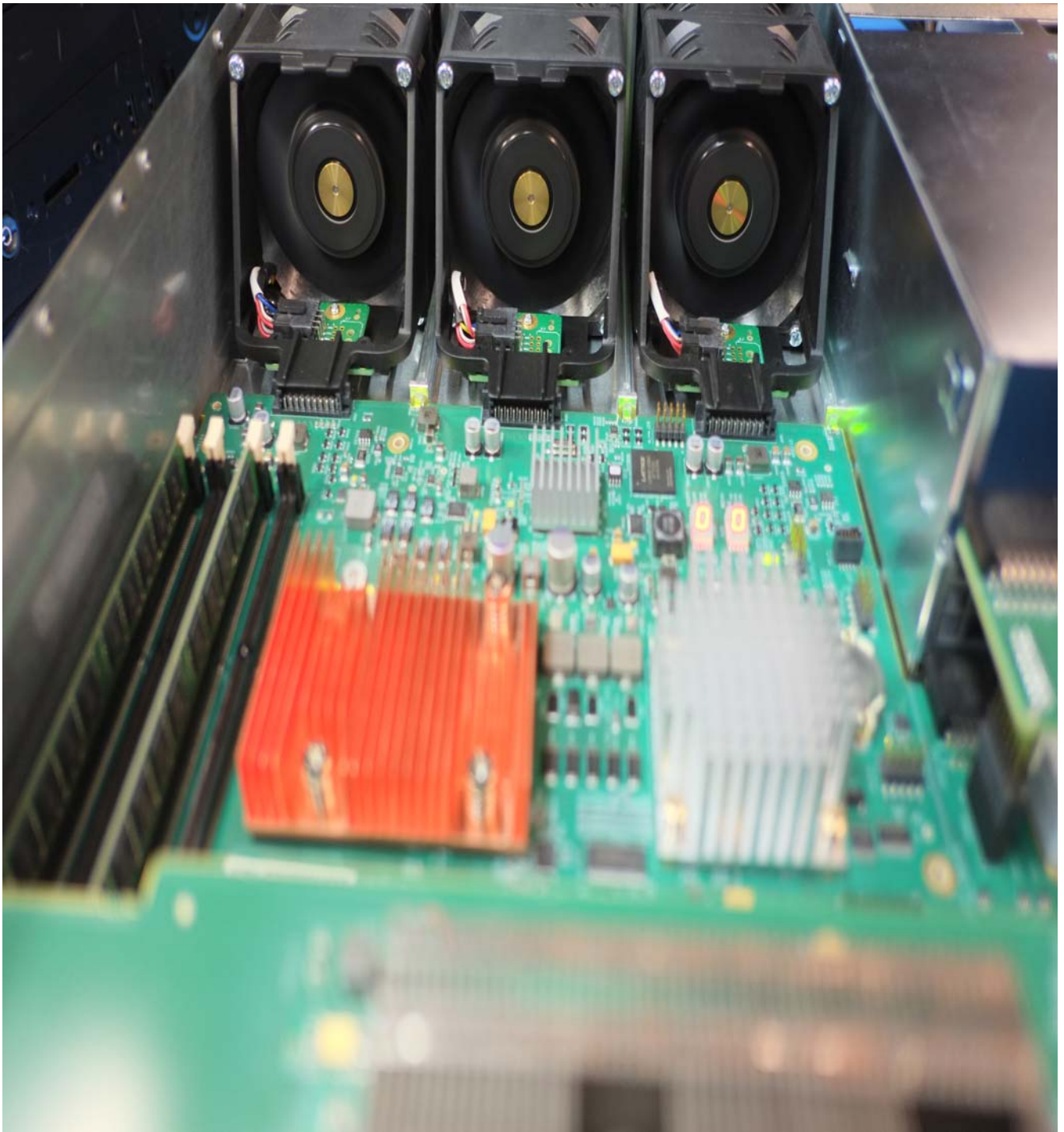
### Cooling

Three twin turbines cool the Juniper QFX10002. Each fan has twin turbines to provide redundancy within the fan itself in case there is a fault. The switch itself has 2 + 1 fan redundancy. The same orange and blue color scheme is used on the fans as well.



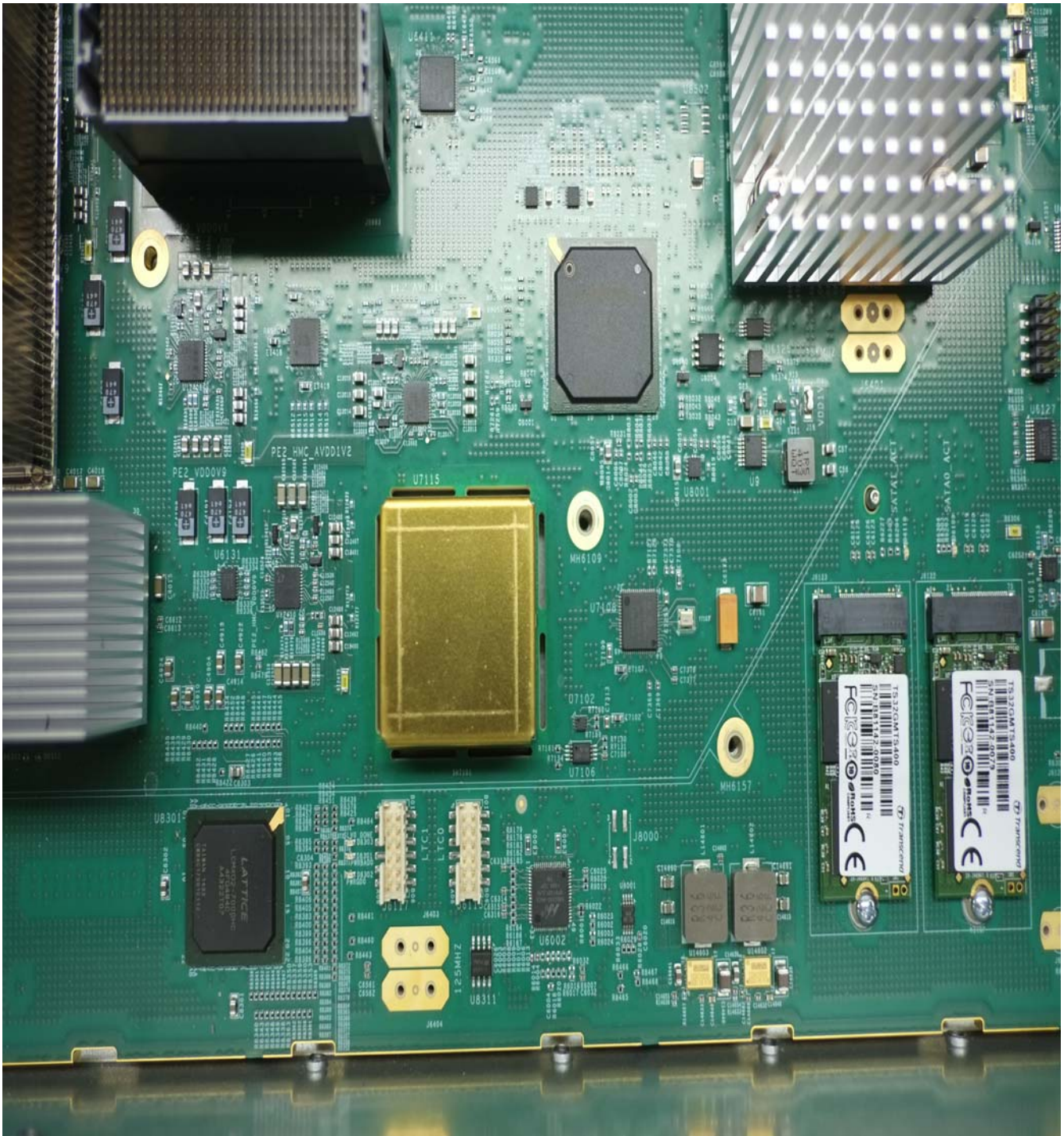


Here's a great show of all three fans installed in the system:



## Precision Timing

The Juniper QFX10002 supports the precision timing protocol (PTP) and can even act as a grandmaster. It has an oven-baked oscillator as shown below:



The primary use cases for PTP are financial trading and time-stamping packets for precision network analytics with Juniper's new Cloud Analytics Engine.

## Linux + Junos

The Juniper QFX10002 boots directly into Linux as soon as you power it on. We have a new software architecture that allows for extreme programmability.

<b>Junos Control Plane Master</b>	<b>Junos Control Plane Backup</b>	PFE Daemon	Platform Daemon	Analytics Daemon	Apache Thrift
VM	VM				
Yocto Linux / KVM					
Control Board: Xeon Processors, DDR3, SSDs					
Juniper Q5 Silicon					

The Juniper QFX10002 supports full in-service software upgrade (ISSU) with two routing engines thanks to the power of Linux KVM and virtualization. The Junos control plane is running inside of two VMs that simulate traditional routing engines in chassis based systems.

Juniper has also provided a hardware abstraction layer (HAL) that runs natively in Linux. This is also true for the platform, analytics, and automation software.

## Junos RPMs and Package Management

The Juniper QFX10002 is bringing big changes to the way you upgrade software. Traditionally Juniper switches have required jinstall packages to upgrade the switch software. These files were large (400MB+) and you had to copy them to the switch and execute CLI commands to upgrade the software.

Wouldn't it be nice if you could simply upgrade a network switch as if it was a Linux server?

We've redesigned everything with the Juniper QFX10002 and have built all of the software around the RedHat Package Management (RPM) system. We have also broken the packages out into the following categories:

- A single Junos control plane with no hardware or platform dependencies.
- Platform specific drivers (interfaces, fans, power supplies, FPGAs).
- Packet forward engine (PFE) drivers (Juniper Q5).

You can now upgrade any of these categories independently from each other. However we have also used the RPM spec file to enforce software dependencies where required. For example if a new version of Junos required a new PFE SDK, we would upgrade both components automatically.

No more having to copy files. You can simply use Juniper's public RPM repository or your own private repository. Simply upgrade the switch with a single command. Please note that Yocto Linux uses the "smart" command instead of "rpm":

```
root@localhost:~# smart update
Loading cache...
Updating cache...
##### [100%]

Fetching information for 'junos smart channel'...
• http://jnpr.net/repo/repomd.xml

repomd.xml
##### [ 50%]
-> http://jnpr.net/repo/primary.xml.gz
primary.xml ##### [ 75%]
-> http://jnpr.net/repo/filelists.xml.gz
filelists.xml
```

##### [100%]

Updating cache...

##### [100%]

Channels have 1 new packages:

qfx-10-f-data-plane-15.1X51-D12.1@x86\_64

Saving cache...

The new RPM tool is in addition to the existing method of “request software add” and doesn’t replace it. You can still use the old method of installing software; however the new RPM method is going to be much easier.

I’m really excited to see RPM integrated into the Juniper QFX10002. You can now use standard Linux tools to upgrade the control plane (Junos), data plane (Juniper Q5), or any of the Linux components (analytics, apache thrift).

You can now use Chef and Puppet to upgrade any of the software on the Juniper QFX10002 just as if it was any other server in your data center.

## Summary

The Juniper QFX10002 is an awesome switch. It packs a big punch in port density in a small 2RU form factor, but at the same time provides massive logical scale, buffer, and an uncompromised feature set. The philosophy behind the Juniper QFX10002 is to simply do more with less. It’s amazing that such a small switch can do so much.

For more information check out these other blogs on the QFX10000 family.

<http://forums.juniper.net/t5/Data-Center-Technologists/QFX10000-a-no-compromise-switching-system-wit...>

<http://forums.juniper.net/t5/Silicon-and-Systems/Powerful-but-Green-How-3-D-Memories-Help-Networking...>

Everyone's Tags: [Juniper Q5 chip](#) [Juniper QFX10002](#) [Juniper switch](#) [View All \(3\)](#)

---

## Comments

---

by [ytti](#) on 03-19-2015 04:24 AM

Interesting stuff. Would love to hear more about how you do lookups on QFX10k with HMC.

How does binary result of bloom filter help finding out egress+rewrite for LPM in IP or exact match in MAC?

---

by  [Doug Hanks \(JNPRdhanks\)](#)  on 03-19-2015 10:39 AM

Hi ytti,

The result of searching for a key in a table is a value, i.e. a nexthop. The nexthop, along with properties of the

egress interface is used for egress processing, including rewrite.

---

by [ytti](#) on 03-19-2015 10:52 AM

Hi Doug,

In meant specifically in relation to bloom filter. As the filter apparently only will tell you 'maybe, no' this is clearly not sufficient to determine the egress port or rewrite.

Just stabbing in the dark here.

So is the bloom filter on-chip, used to pre-classify lookup to single hash-table on HMC? Like perhaps you have bloom filters for every VRF, to figure out which VRF it will be, then bloom filters for every prefix-length to select one of 32 hash-tables? Then you could do hash-lookup on off-chip memory, where you already know VRF and prefix-length?

But I guess prefix-length wouldn't be the best (or rather would be terrible) way to approach this, as then /32 and /24 hash tables would be very large. /2 would be empty, etc.

Generally would love to expansion on the great article, to dig deeper on how Juniper managed to pull this of, what is the lookup process. Really impressed by the HW and liking your SW side work too (rpm with operator repos seems awesome!).

---

by [willygeorge](#) on 03-24-2015 10:13 AM

Hi Doug,


Great write up with a lot of interesting information. Loved your book on the QFX5100 as well. Can you point to a source for more information on Bloom filters? Hopefully somebody will do a similar overview on the PTX platforms as well.

Thanks

Willy

 Posted from Apple iPad

---

by  [bshelton](#) on 04-02-2015 09:00 AM

Doug,



Fantastic blog. One minor note.

You say, "so each ACL consumes 150mW. The Juniper QFX10002 only uses 3mW per firewall entry, which is 98% more efficient." I think your math might be off.

If a typical switch ACL consumes 150mW per firewall entry and the QFX10002 uses 3mW per firewall entry then wouldn't the QFX10002 be 50 times more efficient, and wouldn't that be 2500% (25 times 100%) more efficient?