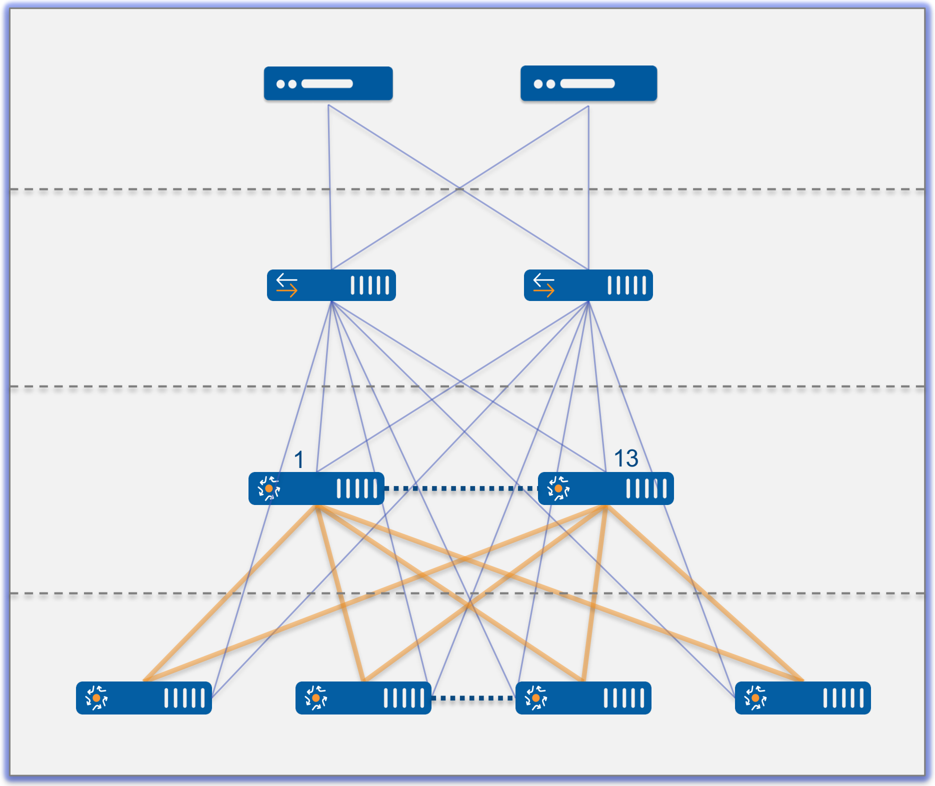
ソフトウェアから見た Disaggregated Distributed Backbone Router (DDBR) へのチャレンジ

Tetsuya Murakami

October 26 2023

DDBR Architecture – Flexibility & High Scalability



Control Plane Cluster
- Commodity multi-core x86 server
- Runs VDR Control & Management planes

Underlay Connector
- Layer 3 fully managed switch
- Out of band (OOB) Control & Management connectivity

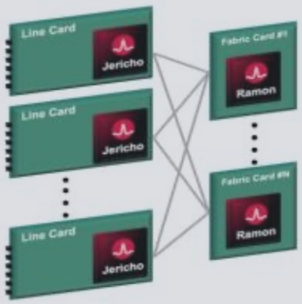
Fabric Card
- Single-chip or Multi-chip Ramon platform
- Future-proof to support any Next-Gen silicon

Line Card
- High Density ports: 10/25/100/400 Gigabit Ethernet
- Jericho 2 today, Jericho 3 or any Next-Gen silicon

Breaking the scale barrier: up to 7,680 x 100GE ports

ARRCUS™

DDBR components – ODM whitebox lineup



BROADCOM

Fabric Card

Ramon – 19.2Tbps

48 x 400GE QSFP-DD Fabric Ports

Ramon3 – 51.2Tbps

64 x 800GE QSFP-DD Fabric Ports

Line Card

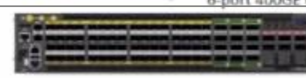
Jericho2

40 x 100GE QSFP28
13 x 400GE QSFP-DD

High Density 100GE Line Card Switch

Jericho2c

64-port 25GE SFP28
12-port 100GE QSFP28
6-port 400GE QSFP-DD Fabric



High Density 25GE/100GE Line Card Switch

Jericho2

10 x 400GE QSFP-DD
13 x 400GE QSFP-DD

400GE Line Card Switch

Jericho2c+

36-port 400GE QSFP-DD
40-port 400GE QSFP-DD Fabric

High Density 400GE Line Card Switch

Jericho3

18-port 800GE QSFP-DD or 36-port 400GE QSFP-DD
20-port 800GE QSFP-DD Fabric

High Density 800GE Line Card Switch

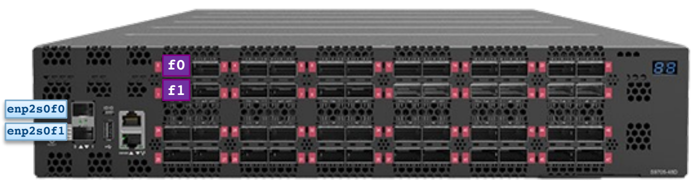
Available today

Available 2024

ARRCUS

ODM hardware – Fabric Card

UfiSpace S9705-48D



- Broadcom Ramon ASIC
- Intel Broadwell-DE processor
- 48x400GE (fabric)
- DDBR's Backplane Switch Fabric

SPECIFICATIONS

PHYSICAL

- 48 x 400GE QSFP-DD fabric ports
- 1 x RJ45 + Micro USB serial console ports
- 1 x 1GBase-T Ethernet port for Out-of-Band management
- 1 x USB 2.0 Type-A general purpose port
- 2 x 10GBase-X SFP+ management ports

Processor	Intel Broadwell-DE 8 core @ 2.0 GHZ
Memory	2 x 32GB DDR4 RDIMM with ECC support
Storage	2 x M.2 SATA SSD 128GB
ASIC	Broadcom Ramon BCM88790
LED	Power status Fan status System status Per port link status Beacon Per fan status Per PSU status
Chassis (WxDxH)	2RU, 436 x 762 x 87.6 mm 17.17 x 30 x 3.45 in. Weight: 24.9 kg (54.9 lbs)
Redundancy	Hot swappable, 1+1 redundant PSU Hot swappable, 3+1 redundant fans

ENVIRONMENTAL

Power Supply	AC input: 200 ~ 240V, 12.5A, 50 ~ 60Hz DC input: -40 ~ -72V, 60 ~ 33A Typical power: 980 Watts (no transceiver)
Max. Operating Specs.	Operating temperature: 0°C to 45°C (32°F to 113°F) Operating humidity: 5% to 85% (RH), noncondensing ltitude: 1,829 m (6,000 ft.)
Max. Non-Operating Specs.	Storage temperature: -40°C to 70°C (-40°F to 158°F) Storage humidity: 5% to 95% (RH), non-condensing

PERFORMANCE

Switching Capacity	4.8 Tbps
Packet Throughput	8 billion cells

REGULATORY COMPLIANCE

Safety	NEBS Level 3 UL 62368-1 IEC/EN 60950-1 IEC/EN 62368-1 BSMI CNS 14336-1	EMC	NEBS Level 3 FCC Part 15, Subpart B, Class A EN 55032, Class A EN 300 386 EN 55024 EN 55035 BSMI (CNS 13438), Class A VCCI-CISPR 32:2016, Class A VCCI 32-1:2016, Class A

Specifications are subject to change without notice.

ARRCUS™

ODM hardware – Line Card or Standalone

UfiSpace S9700-53DX

High Density 100GE

- Broadcom J2 ASIC
- Intel Broadwell-DE processor
- Access ports: 40x100GE
  - or 80x10/25GE w/ breakout
- Fabric ports: 13x400GE
- Switching capacity **4.8Tbps**
- Large routing table
- 8GB deep packet buffer
- Redundant PS and fans

SPECIFICATIONS

PHYSICAL

- 40 x 100GE QSFP28 service ports
- 13 x 400GE QSFP-DD fabric ports
- 1 x RJ45 + Micro USB serial console ports
- 1 x 1GBase-T Ethernet port for Out-of-Band management
- 1 x USB 2.0 Type-A general purpose port
- 2 x 10GBase-X SFP+ management ports

Processor	Intel Broadwell-DE 8 core @ 2.0 GHZ
Memory	2 x 32GB DDR4 RDIMM with ECC support
Storage	2 x M.2 SATA SSD 128GB
ASIC	Broadcom Jericho2 BCM88690
LED	Power status Fan status System status Per port link status Beacon Per fan status Per PSU status
Chassis (WxDxH)	2RU, 436 x 762 x 87.6 mm 17.17 x 30 x 3.45 in. Weight: 26.7 kg (58.9 lbs)
Redundancy	Hot swappable, 1+1 redundant PSU Hot swappable, 3+1 redundant fans

ENVIRONMENTAL

Power Supply	AC input: 200 ~ 240V, 12.5A, 50 ~ 60Hz DC input: -40 ~ -72V, 60 ~ 33A Typical power: 750 Watts (no transceiver)
Max. Operating Specs.	Operating temperature: 0°C to 45°C (32°F to 113°F) Operating humidity: 5% to 85% (RH), noncondensing ltitude: 1,829 m (6,000 ft.)
Max. Non-Operating Specs.	Storage temperature: -40°C to 70°C (-40°F to 158°F) Storage humidity: 5% to 95% (RH), non-condensing

PERFORMANCE

Switching Capacity	4.8 Tbps
Packet Throughput	2000 Mpps

REGULATORY COMPLIANCE

Safety	NEBS Level 3 UL 62368-1 IEC/EN 60950-1 IEC/EN 62368-1 BSMI CNS 14336-1	EMC	NEBS Level 3 FCC Part 15, Subpart B, Class A EN 55032, Class A EN 300 386 EN 55024 EN 55035 BSMI (CNS 13438), Class A VCCI-CISPR 32:2016, Class A VCCI 32-1:2016, Class A

Specifications are subject to change without notice.

ARRCUS

ODM hardware – Line Card

UfiSpace S9700-23D

400GE Line Card

- Broadcom J2 ASIC
- Intel Broadwell-DE processor
- Access ports: 10x400GE
  - or 40x100GE w/ breakout
- Fabric ports: 13x400GE
- 400GE ZR support
- Switching capacity **4.8Tbps**
- Large routing table
- 8GB deep packet buffer
- Redundant PS and fans

ARRCUS

SPECIFICATIONS

PHYSICAL

- 10 x 100/400G QSFP-DD service ports
- 13 x 400G QSFP-DD fabric ports
- 1 x RJ45 & Micro USB serial console ports
- 2 x 10GBase-X SFP+ management ports
- 1 x 100/1000M RJ45 management port
- 1 x USB 2.0 Type-A port

Processor	Intel Broadwell-DE 8-Core @ 2.0GHz
Memory	64GB DDR4
Storage	256GB SSD
ASIC	Broadcom Jericho2 BCM88690 Broadcom OP2 BCM16K
BMC	AST2400
LED	Power status Fan status System status Per port link status Beacon Per fan status Per PSU status
Chassis (WxDxH)	2RU, 436 x 762 x 87.7 mm or 17.17" x 30" x 3.45" Weight: 18.77kg or 41.38lb
Redundancy	Hot swappable, 1+1 redundant PSU Hot swappable, 3+1 redundant Fans

ENVIRONMENTAL

Power Specs.	AC input: 200 to 240V, 12.5A DC input: -40 to -72V, 60A Typical power: 283 Watts (no transceiver)
Max. Operating Specs.	Operating temperature: 0°C to 45°C (32°F to 113°F) Operating humidity: 5% to 85% (RH), non-condensing
Max. Non-Operating Specs.	Storage temperature: -40°C to 70°C (-40°F to 158°F) Storage humidity: 5% to 93% (RH), non-condensing

PERFORMANCE

Switching Capacity	4.8Tbps
Deep Buffer	8GB

REGULATORY COMPLIANCE

Safety	UL 62368-1 IEC 62368-1 BSMI NOM	EMC	FCC Part 15, Subpart B, Class A ICES-003, Class A EN 55032, Class A EN 55024 EN 55035 EN 62479 EN 50663 EN 300 386 EN 301 489 EN 303 413 BSMI VCCI CISPR 32, Class A AS/NZS CISPR 32, Class A ANATEL NEBS GR-1089, NEBS Level 3
Environment	WEEE RoHS GR-63, NEBS Level 3		

Specifications are subject to change without notice.

ODM hardware – Line Card

UfiSpace S9710-76D

High Density 400GE Line Card

- Broadcom J2c+ ASIC
- Intel Skylake-D processor
- Access ports: 36x40/100/400GE
- Fabric ports: 40x400GE
- 400GE ZR and OpenZR+ support
- Switching capacity 1**4.4Tbps**
- Large routing table
- 16GB deep packet buffer
- SyncE and IEEE1588v2
- Redundant PS and fans

ARRCUS

SPECIFICATIONS

PHYSICAL

- 36 x 40/100/400G QSFP-DD service ports supporting 400ZR and OpenZR+
- 40 x 400G QSFP-DD fabric ports
- 1 x RJ45 & Micro USB serial console ports
- 2 x 10GBase SFP+ management ports
- 1 x 100/1000M RJ45 management port
- 1 x USB 3.0 Type-A port

Processor	Intel Skylake-D 8-Core @ 1.9GHz
Memory	64GB DDR4
Storage	256GB SSD
ASIC*	Broadcom Jericho2c+ BCM88850 Broadcom OP2 BCM16K (Premium)
BMC	AST2400
Timing Interfaces	1 x 10MHz input/output SMB 1 x 1PPS input/output SMB
Timing Support	Stratum 3E OCXO ITU-T Synchronous Ethernet (SyncE) IEEE 1588v2 (Default Profile, G.8265.1 G8275.1, G.8275.2 profiles) T-TC, T-BC/OC
Chassis (WxDxH)	2 RU, 436 x 762 x 87.7 mm or 17.17" x 30" x 3.45" Weight: 26.95kg or 59.41lb
Redundancy	Hot-swappable, 1+1 Redundant PSU Hot-swappable, 3+1 Redundant fans

*OP2 available for premium SKU only

ENVIRONMENTAL

Power Specs.	AC input: 200 to 240V, 16A DC input: -40 to -72V, 80A Typical power: 667 Watts (no transceiver)
Max. Operating Specs.	Operating temperature: 0°C to 45°C (32°F to 113°F) Operating humidity: 5% to 85% (RH), non-condensing
Max. Non-Operating Specs.	Storage temperature: –40°C to 70°C (–40°F to 158°F) Storage humidity: 5% to 93% (RH), non-condensing

PERFORMANCE

Switching Capacity	14.4Tbps
Deep Buffer	16GB

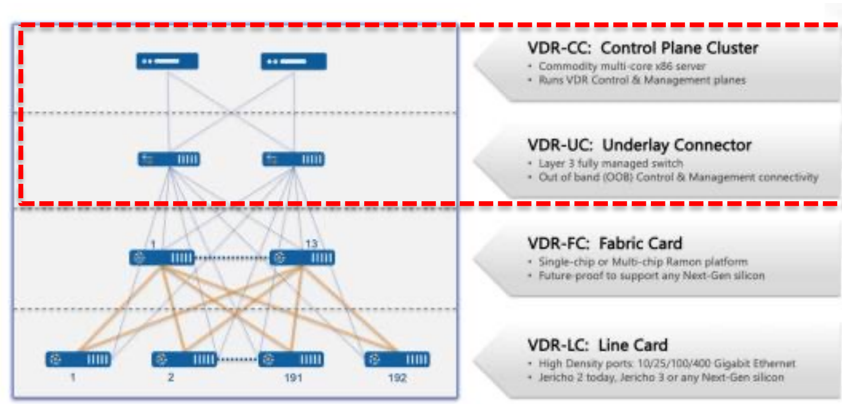
REGULATORY COMPLIANCE

Safety	UL 62368-1 IEC 62368-1 BSMI CNS 15598-1	EMC	FCC Part 15B, Subpart B, Class A ICES-003, Class A EN 55032, Class A EN 55024 EN 55035 EN 62479 EN 50663 EN 300 386 EN 301 489 EN 303 413 BSMI CNS 15598-1 VCCI CISPR 32, Class A
Environment	WEEE RoHS		

Specifications are subject to change without notice.

Compute Server & Underlay Connector

- CS (Compute Server)
  - x86 Server 20-core CPU, 128GB RAM
  - 2x NVMe SSD RAID 1 storage
  - NIC:
    - Intel X710 10GE NIC
    - Intel E810 100GE NIC
    - Mellanox ConnectX-6 DX



- UC (Underlay Connector):
  - Broadcom Trident3 switch
  - Quanta IX8A

ARRCUS™

DDBR is ready?

- Hardware design is fully defined.

However,

- Software design/architecture is very unclear.
    - No available open source – Can’t try DDBR easily.
    - No suitable Operating System.
    - Need to design/develop software from the scratch.
    - There are several challenging….

Slot identity and interface naming?

Problem:

- Unlike a real chassis, DDBR does not have physical slot identifiers that can be used in interface naming.
- How do we identify and name interfaces on DDBR LCs.

Solution:

- User configures a slot number against the serial number for a LC/FC to make it a part of the cluster
- Until this happens LCs and FCs cannot join the cluster and participate in the data path
- Examples:
  - eth5_20 ==> Port 20 on slot 5
  - eth5_20s3 ==> Break out #3 on port 20 on slot 5
  - eth5_20s3_1 ==> Sub-interface 1 on break out #3 on port 20 on slot 5

ifindex allocation problem?

Problem:

- ifindex is allocated by Linux kernel when creating interfaces in kernel
- ifindex is used system-wide to uniquely identify an interface
- In DDBR, there are many Linux kernel instances and interfaces are distributed across many LCs
- How do we ensure a unique system-wide ifindex

Solution:

- Move ifindex allocation responsibility to interface manager outside of Linux
- Change ASIC interface handling to accept app provided ifindex
- Entails major rework of interface information flow within the system

Interface discovery and punt path?

Problem:

- How does NOS control plane get to know about interfaces located on LCs
- How does control plane send and receive packets through those interfaces

Solution:

- As a LC joins the cluster its interfaces are discovered, virtual representations of those interfaces are created on the CS
- A L3 tunnel (i.e., VxLAN) is provisioned between the CS and each LC to transmit and receive packets over these interfaces

How distributing software components?

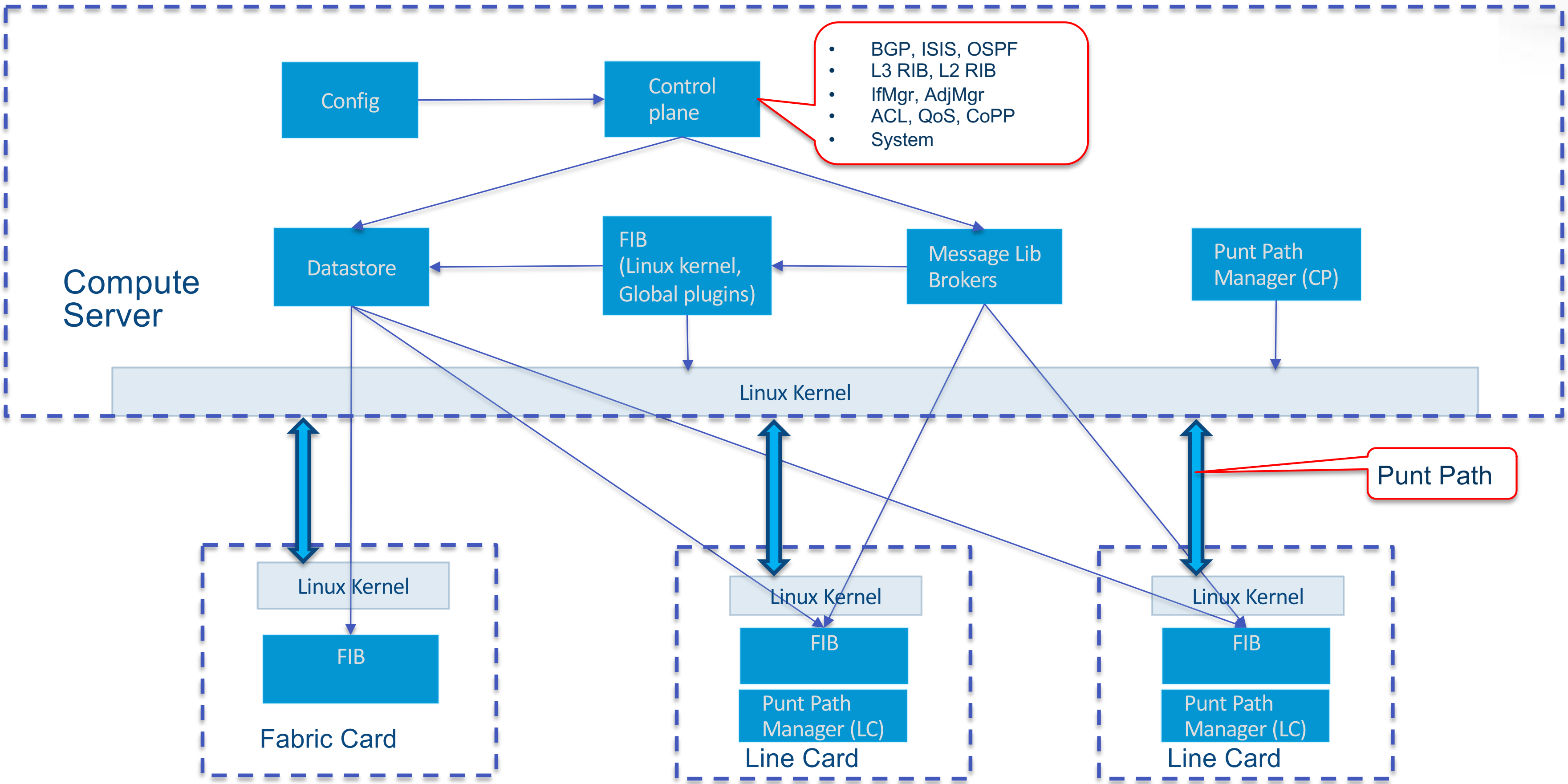
Problem:

- What software components need to be distributed to LC/FC?
- Does the operating system (i.e, Linux) need to be synced up between LC/FC/CS? If yes, how to sync up?

Solution:

- In order to simplify the software architecture, only FIB should be distributed to LC/FC
- FIB on CS should manage the entire information and distribute partial information to FIB on LC/FC
- The operating system on CS/LC needs to be synced up because FIB information is needed to handle interfaces, ICMP error, etc…
- The operating system running on CS has the entire FIB. The operating system running on LC has the partial FIB only.

Distributed software architecture



Global identifier problem?

Problem:

- Certain data path identifiers required by ASIC need to be globally scoped.
- ASIC specific
- Traditional solution of piggy backing on top of control plane download sucks
- How do we allocate and distribute such globally scoped, system-wide ASIC specific identifiers

Solution:

- A global FIB that runs on the CS allocates the global identifier.
- A global identifier is distributed to FIB instances running on LC.
- ASIC can accept it?

Runtime orchestration?

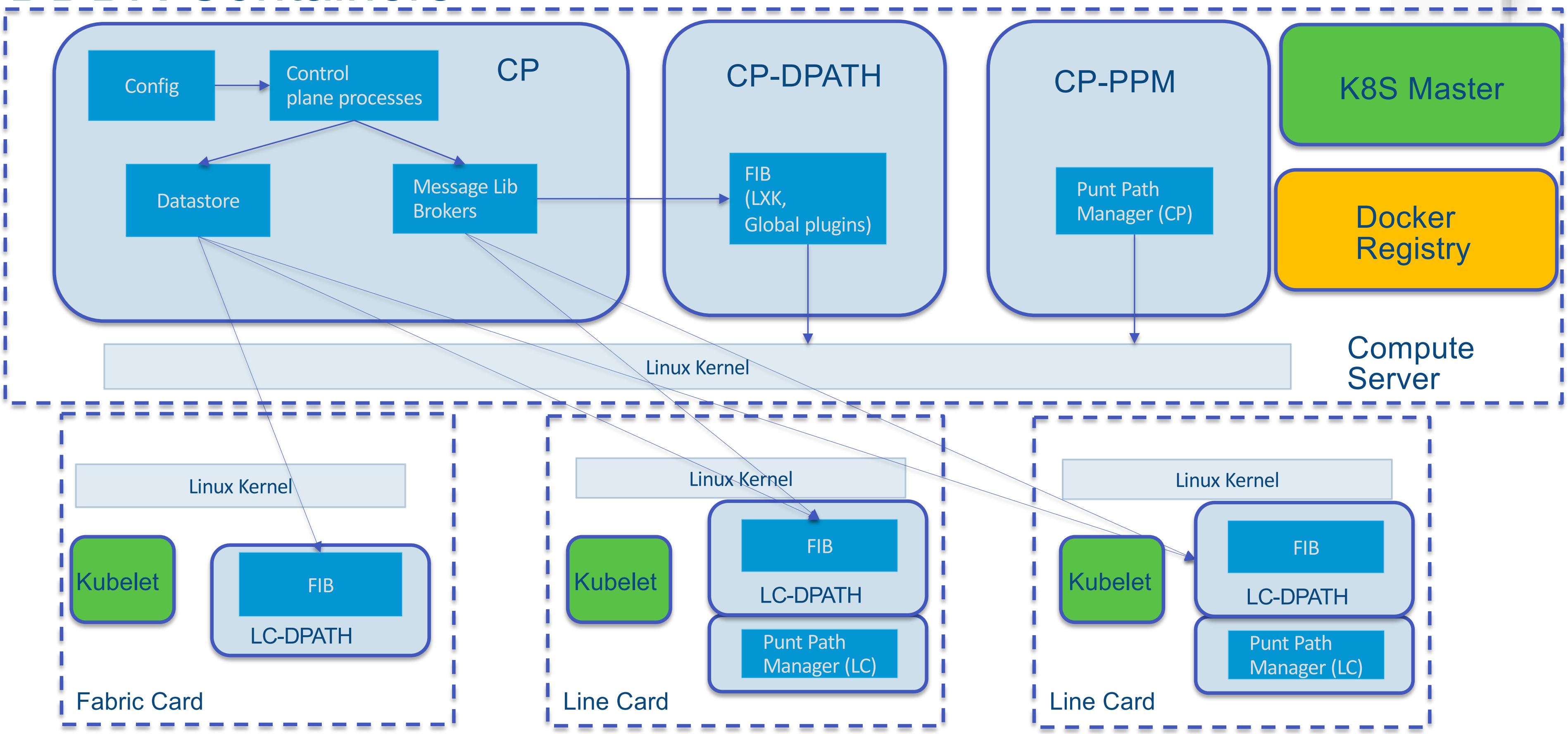
Problem:

- How do we download and run the right set of NOS binaries on the appropriate nodes
- How do we run an active and a standby instance of the control plane on the same host to get soft high availability

Solution:

- Break up NOS into several different container images
- Containers are lightweight and provide enough separation to simulate an entire chassis on a single host
- Choice of several open source orchestration systems to manage and interconnect containers
- Kubernetes (k8s) might be chosen as the orchestration system

DDBR Containers



DDBR Underlay?

Problem:

- ALL nodes need a host OS
- Who does the installation of the host OS
- Who configures the Control Plane Switch (CPS)
- Who assigns node addresses in the control plane network

Solution:

- DDBR underlay tools – collection of tools and scripts that can bootstrap a DDBR cluster from bare metal
- Installs host OS on all nodes, configures the CPS and the node interfaces that connect to it
- Creates ssh keys and login on all nodes for internal access
- Runs NTP between nodes
- Runs an internal DHCP, DNS server, Docker registry service and finally starts off the k8s orchestration services

Networking functions supported in the operating system

Problem:

- How to leverage the networking functions in Linux like bridge, lacp, etc
- How to manage the bridge interface among multiple LCs.
- How to manage lacp on LC and/or CS.

Solution:

- Create the bridge interfaces on multiple LCs and connecting to the path/interface toward to FC.
- Disable lacp on LC and running on lacp on CS only.
- The lacp packets is managed by the punt path manager.
  - It causes some overhead to process lacp packets. Hence, it is difficult to achieve lacp fast.

ARRCUS

Summary

- Hardware design for DDBR is securely defined.
- Software design/architecture for DDBR has not been discussed so far…
  - Need to consider a lot of stuff related to software architecture for DDBR
  - Require TRY & ERROR approach to verify the architecture.
  - Take long time for the development
- No open source tools, No operating system suitable for DDBR
  - Need to modify/tweak the existing open source, operating system.

Thank You!