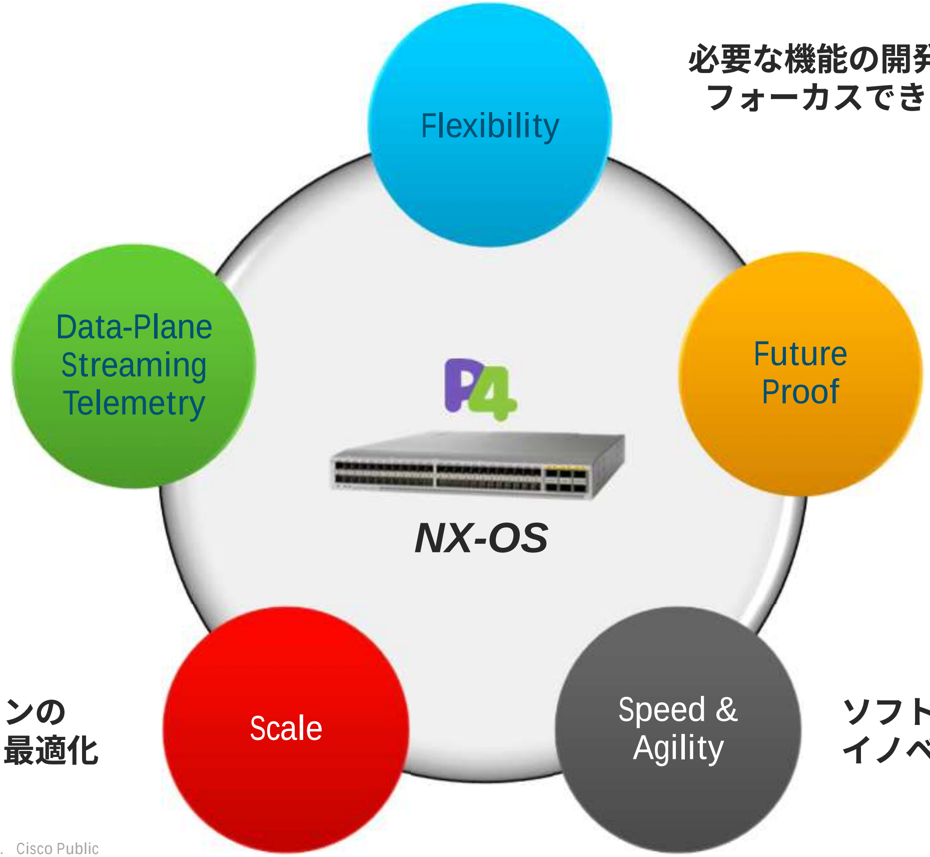
CiscoにおけるP4の活用と展望

シスコシステムズ合同会社
テクニカルソリューションアーキテクト
佐藤 哲大

プログラマブルスイッチ



必要な機能の開発に
フォーカスできる

データプレーンの
詳細情報を取得できる

機器の買い替えなしに
新しい機能を提供できる

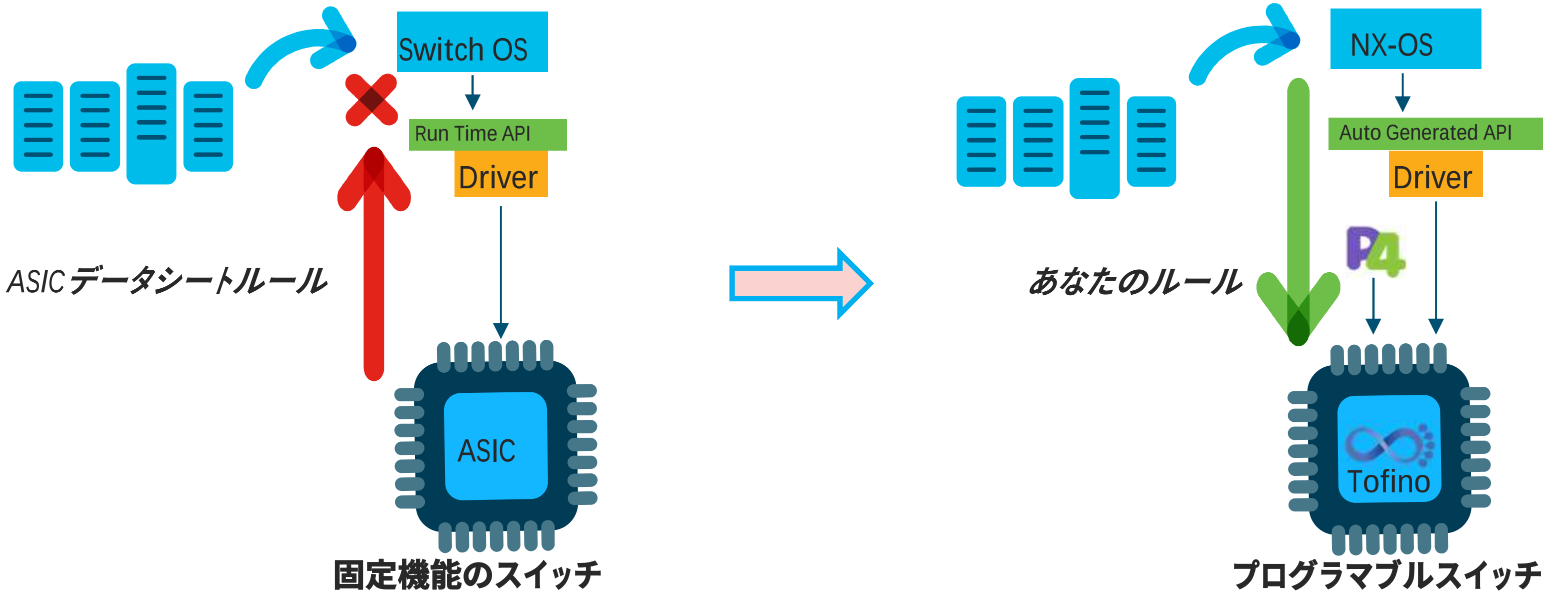
データプレーンの
リソース配分を最適化

ソフトウェアのスピードで
イノベーションを起こせる

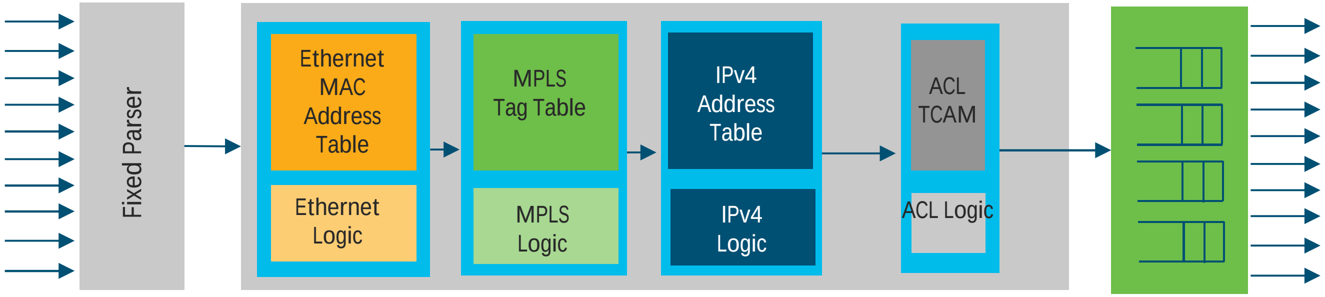
ネットワークデザイン今昔

(昔ながらの) ボトムダウンデザイン

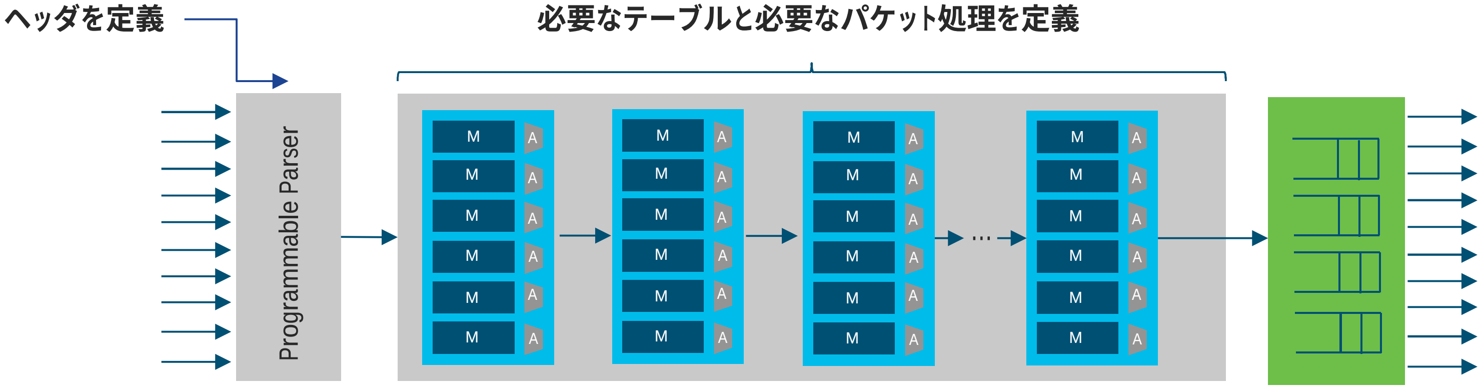
トップダウンデザイン



固定 vs. プログラマブルパケット処理



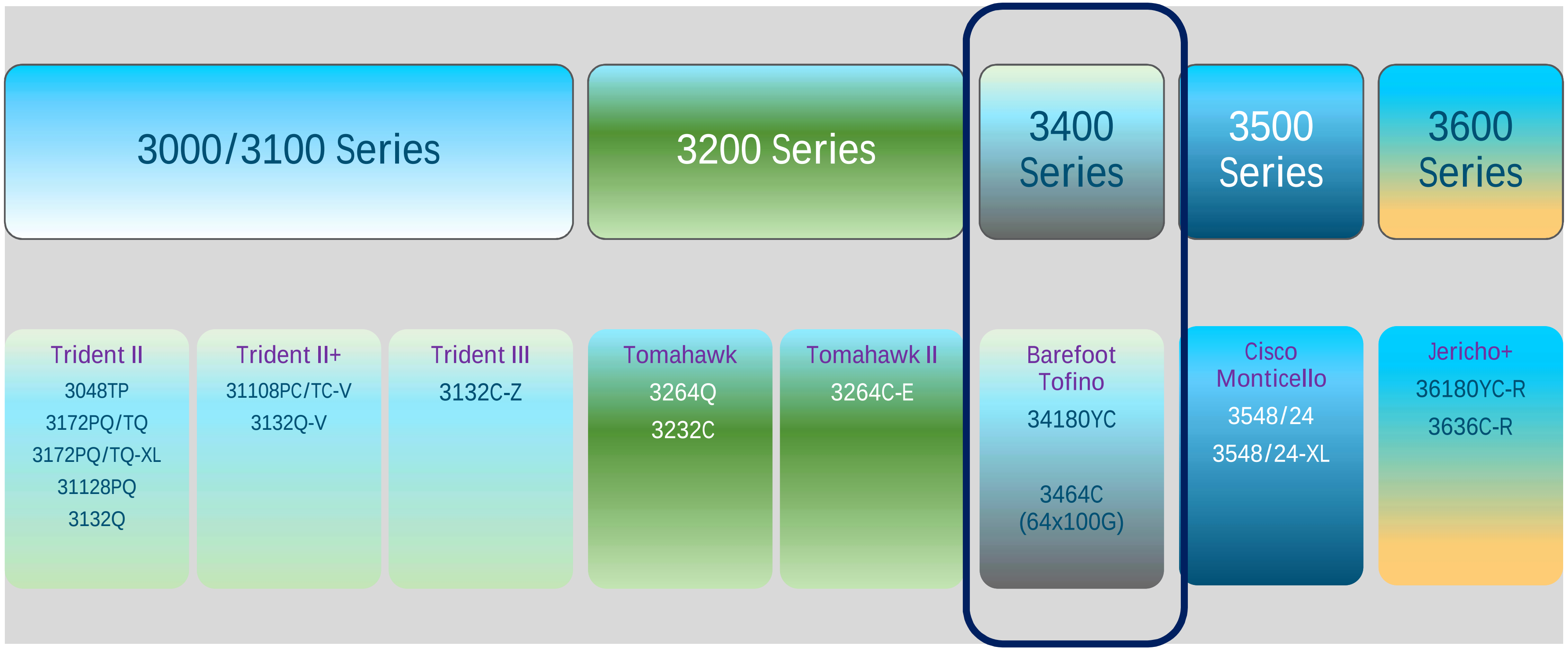
固定パイプライン : 機能とテーブルサイズはASICの設計時に固定



プログラマブルパイプライン: ユーザーがMatch-Actionロジックを定義

プロダクト

Nexus 3000ポートフォリオ



3000/3100 Series

3200 Series

3400 Series

3500 Series

3600 Series

Trident II
3048TP
3172PQ/TQ
3172PQ/TQ-XL
31128PQ
3132Q

Trident II+
31108PC/TC-V
3132Q-V

Trident III
3132C-Z

Tomahawk
3264Q
3232C

Tomahawk II
3264C-E

Barefoot Tofino
34180YC
3464C (64x100G)

Cisco Monticello
3548/24
3548/24-XL

Jericho+
36180YC-R
3636C-R

Nexus 3400 Series

34180YC

48p 10/25G SFP28 + 6p 40/100G QSFP28

Barefoot Tofino 1.8T ASIC

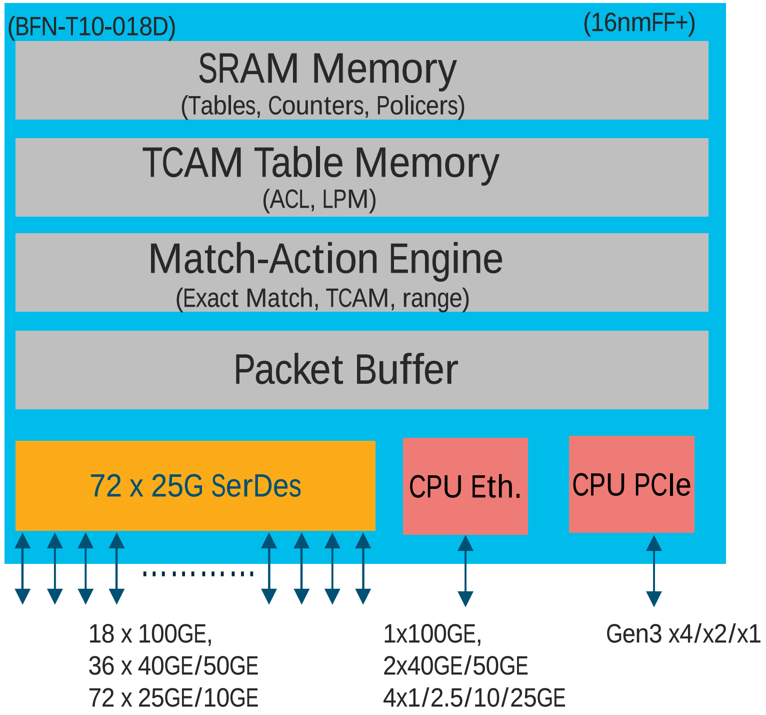
3464C

64p 40/100G QSFP28

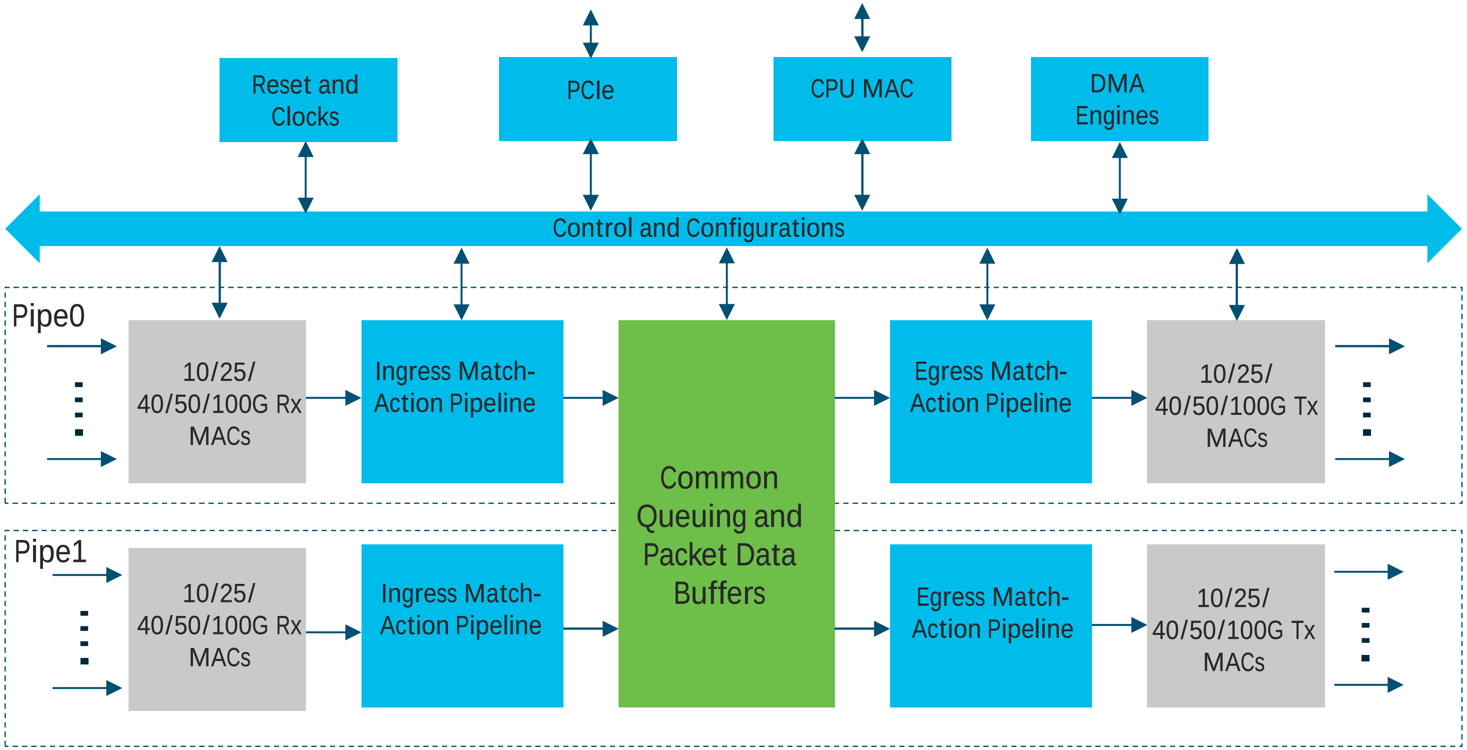
Barefoot Tofino 6.4T ASIC

Barefoot Tofino 1.8T ASIC Architecture

- BFN-T10-018D from Tofino family
- 1.8Tbps Single Chip Ethernet Switch
- 2 Pipes @0.9 Tbps
- P4-programmable pipeline
- Single 20 MB Unified Packet Buffer
- Customizable Low latency

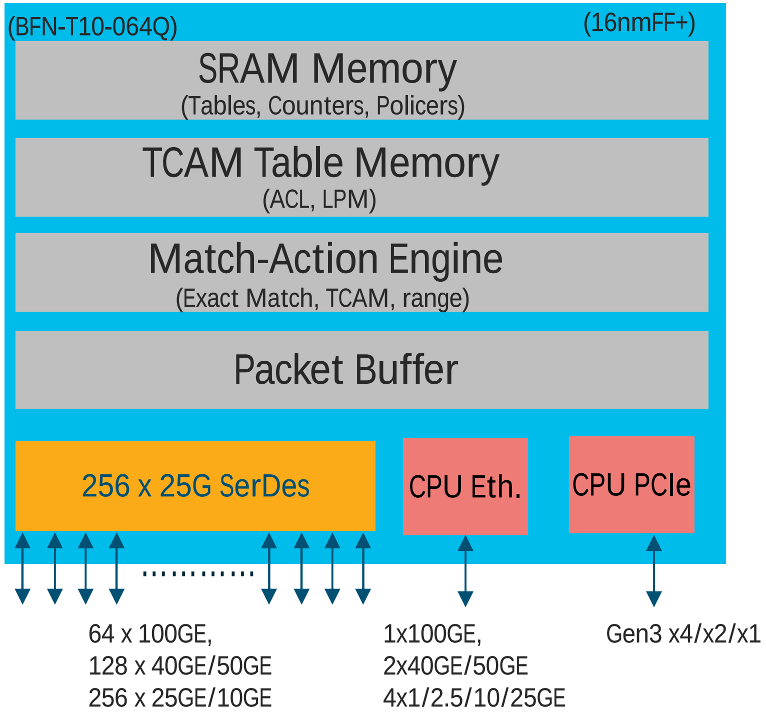


Tofino 1.8T dual-pipeline Block Diagram

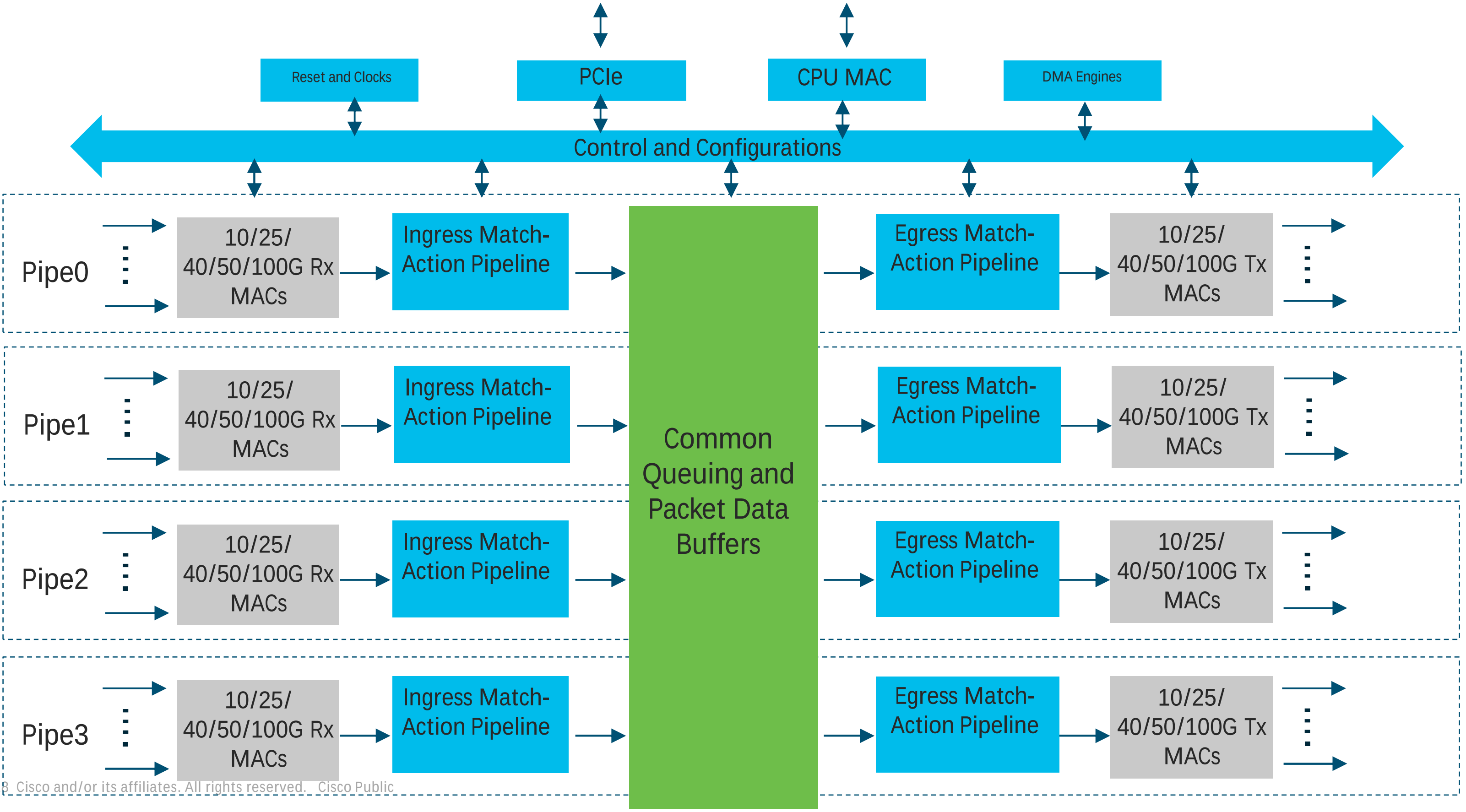

Reset and Clocks
PCIe
CPU MAC
DMA Engines
Control and Configurations
Pipe0
10/25/ 40/50/100G Rx MACs
Ingress Match-Action Pipeline
Common Queuing and Packet Data Buffers
Egress Match-Action Pipeline
10/25/ 40/50/100G Tx MACs
Pipe1
10/25/ 40/50/100G Rx MACs
Ingress Match-Action Pipeline
Egress Match-Action Pipeline
10/25/ 40/50/100G Tx MACs

Barefoot Tofino 6.4T ASIC Architecture

- BFN-T10-064Q from Tofino family
- 6.4 Tbps Single Chip Ethernet Switch
- 4 Pipes @1.6 Tbps
- P4-programmable pipeline
- Single 22 MB Unified Packet Buffer
- Customizable Low latency



Tofino 6.4T quad-pipeline Block Diagram

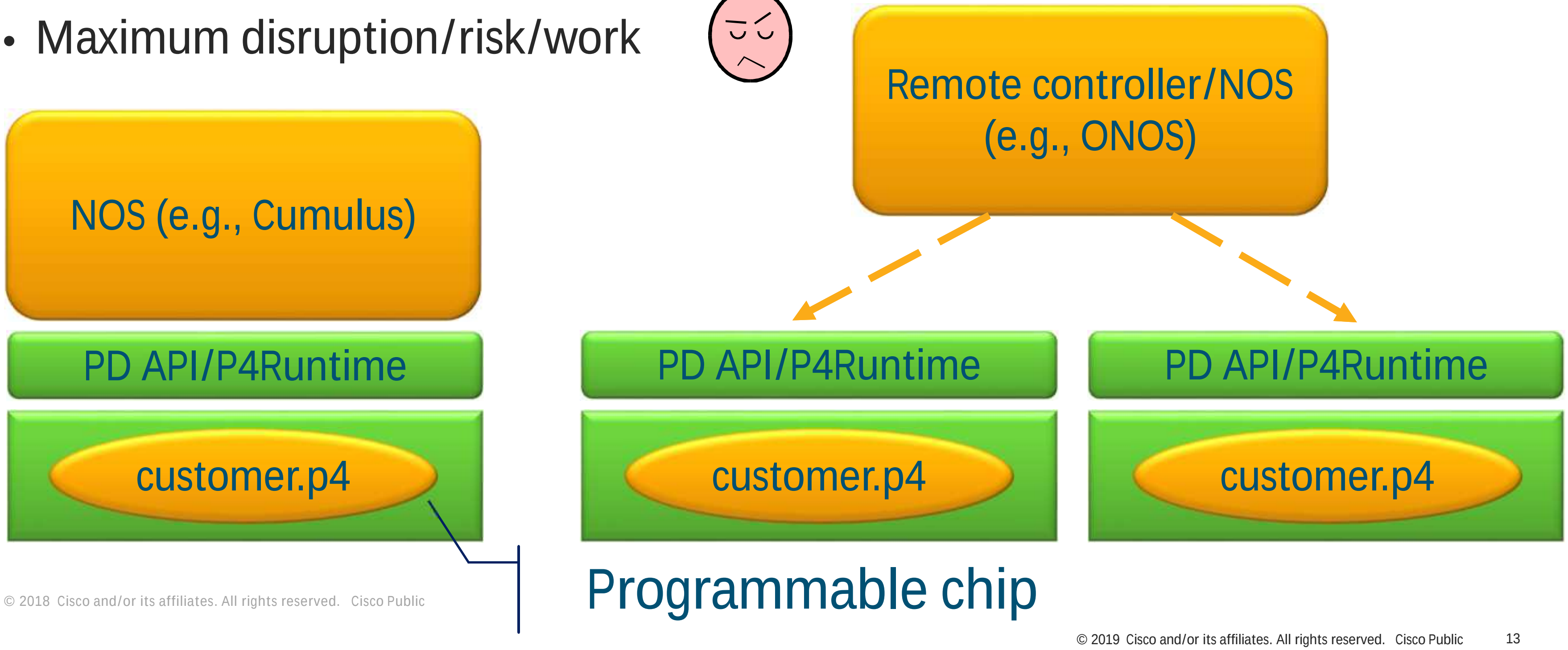

Reset and Clocks
PCIe
CPU MAC
DMA Engines
Control and Configurations
Pipe0
10/25/ 40/50/100G Rx MACs
Ingress Match-Action Pipeline
Common Queuing and Packet Data Buffers
Egress Match-Action Pipeline
10/25/ 40/50/100G Tx MACs
Pipe1
10/25/ 40/50/100G Rx MACs
Ingress Match-Action Pipeline
Egress Match-Action Pipeline
10/25/ 40/50/100G Tx MACs
Pipe2
10/25/ 40/50/100G Rx MACs
Ingress Match-Action Pipeline
Egress Match-Action Pipeline
10/25/ 40/50/100G Tx MACs
Pipe3
10/25/ 40/50/100G Rx MACs
Ingress Match-Action Pipeline
Egress Match-Action Pipeline
10/25/ 40/50/100G Tx MACs

利用モデル

ホワイトボックスモデル

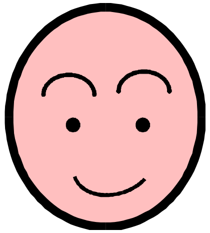
- Maximum flexibility
- Maximum disruption/risk/work

🟦 Platform vendor (Cisco)
🟩 Chip vendor (Barefoot)
🟧 Customer/open source



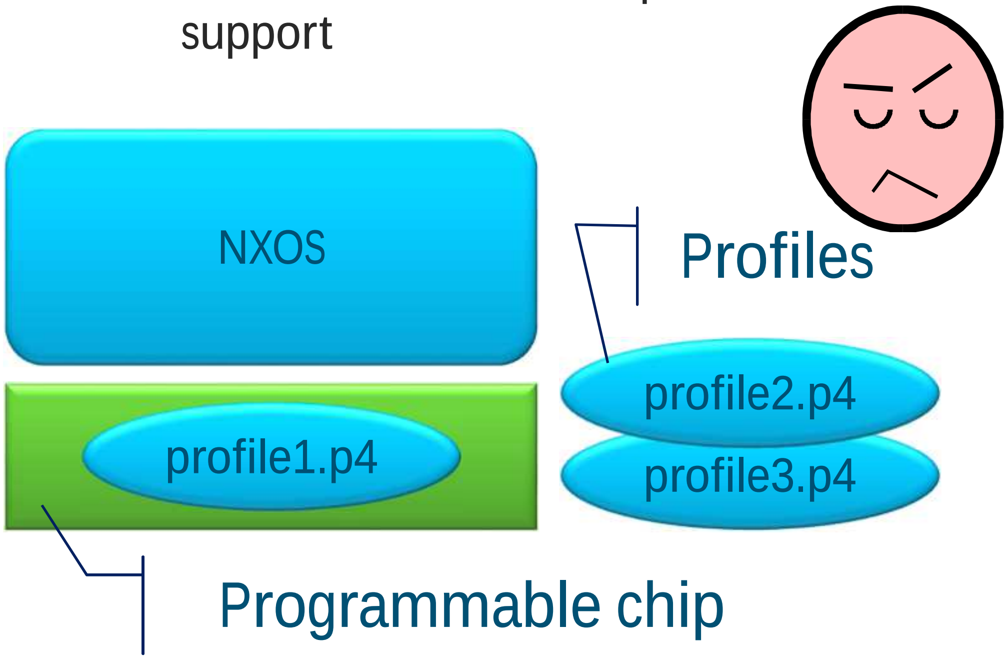
ターンキーモデル

- Deployment as usual
  - Familiar features and interfaces
- Resource optimization
- Future proof
- Feature agility
- Streaming telemetry

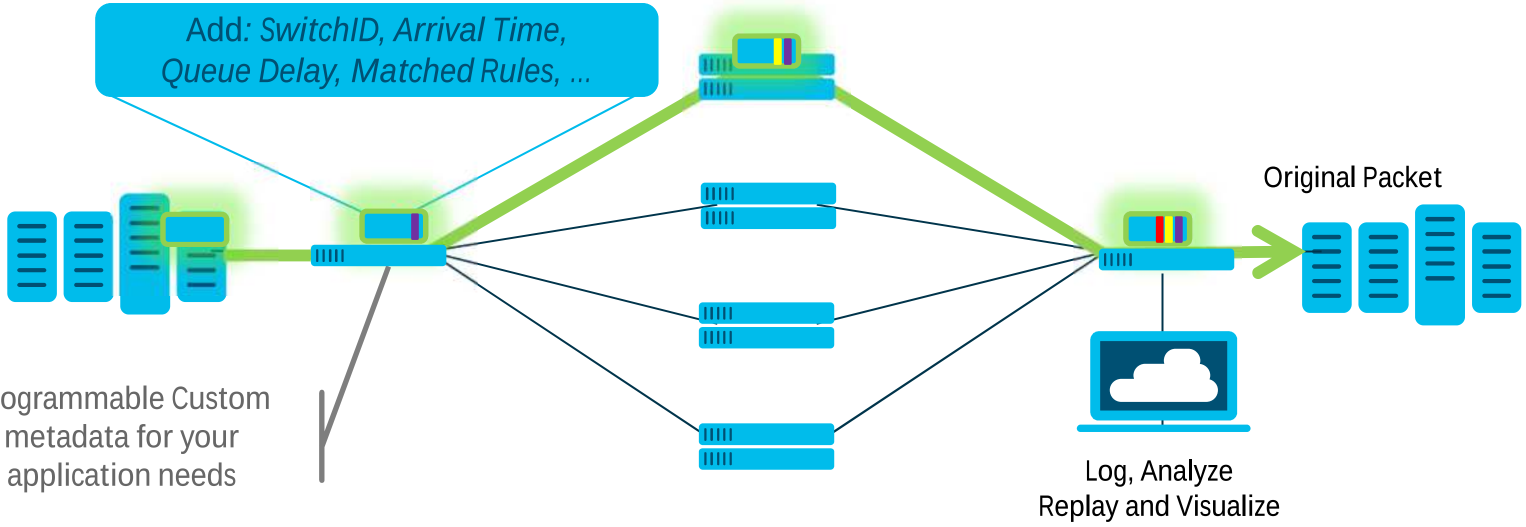


🟦 Platform vendor (Cisco)
🟩 Chip vendor (Barefoot)
🟧 Customer/open source

- No flexibility
  - No custom feature and protocol support



Inband Network Telemetry (INT)



Add: *SwitchID, Arrival Time, Queue Delay, Matched Rules, …*

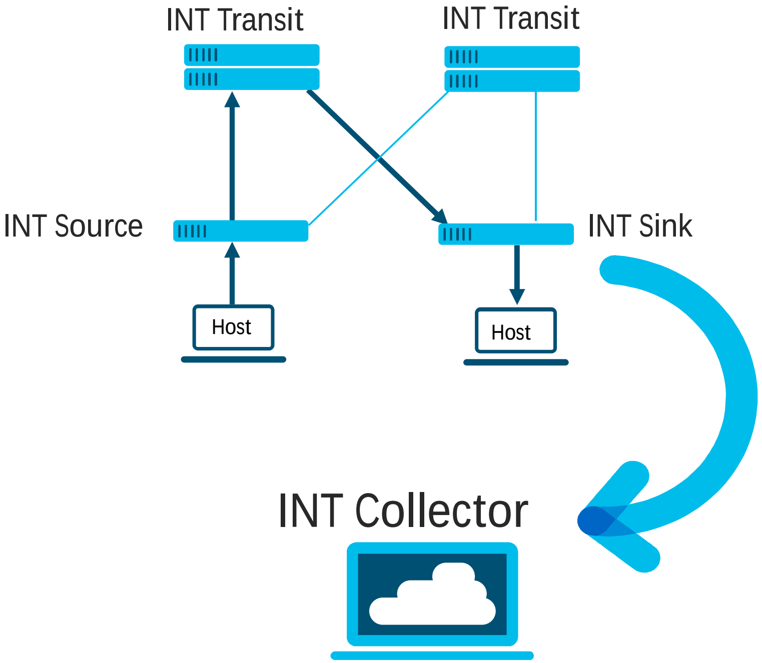
Programmable Custom metadata for your application needs

Original Packet

Log, Analyze Replay and Visualize

Inband Network Telemetry (INT)

- INT Sourceノードにおいて、データパケットに最初のレコード (INT命令とメタデータ) を挿入
- INT Transit ノードにおいて2つ目のレコード (メタデータ) を追加
- INT Sinkノードにおいて、3つ目のレコードを追加
- INT Sink ノードはINT レコードを削除して INT コレクタに送信、オリジナルのパケットはそのまま転送



スイッチ毎に取得可能な情報

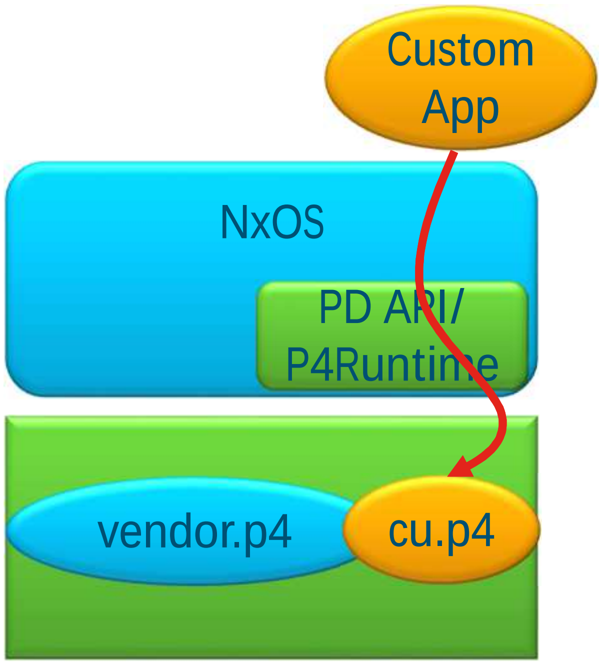
Flow Watch List (zoom-in view per 5-tuple of flow + DSCP bits) – 2K	Drop Watch List (Drop due to various drop reasons) - 512	Queue Watch List (queue depth or latency exceeds configured threshold)
Switch ID	Switch ID	Switch ID
Hop latency	Ingress Port ID	Hop latency
Queue ID + Queue occupancy	Egress Port ID	Queue ID + Queue occupancy
Ingress timestamp	Queue ID	
Egress timestamp	Drop Reason	

- Node-to-Node: Reserved DSCP bit will be inserted temporarily in data packets to indicate that packets also carry INT data
- Node-to-Collector: A UDP encapsulation is used to pack collected INT stack at INT Sink and send to collector. Flow-affinity is maintained to send same flow-record to same collector for easy processing

ハイブリッドモデル

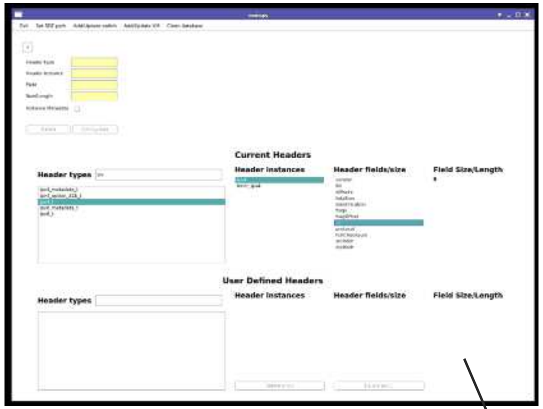
- 🟦 Platform vendor (Cisco)
- 🟩 Chip vendor (Barefoot)
- 🟧 Customer/open source/

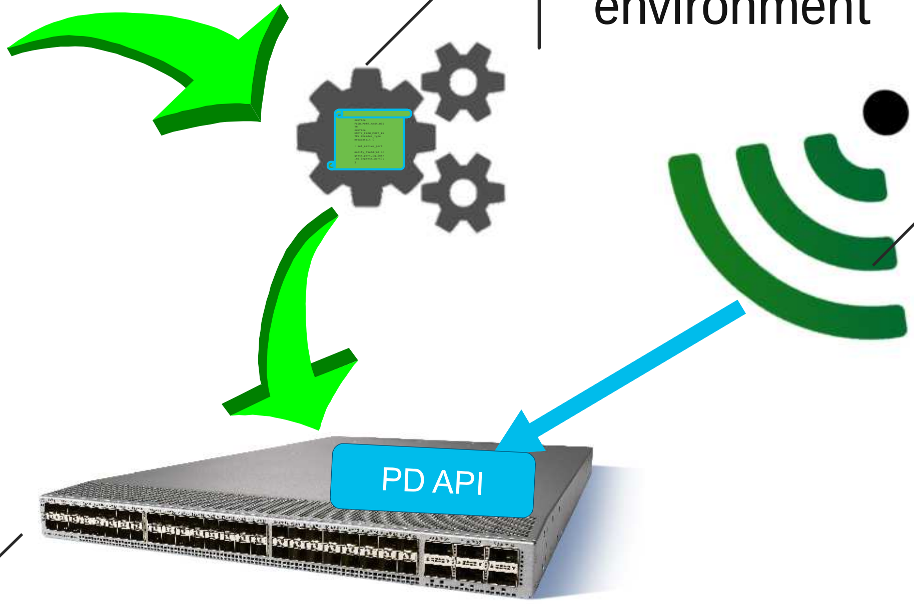
- Best of breed
- Deployment as usual
  - Familiar features and interfaces
- Minimum development effort
  - Leverage existing functions in building new features



Minimize disruption and risk!

data Plane Incremental Programming Environment





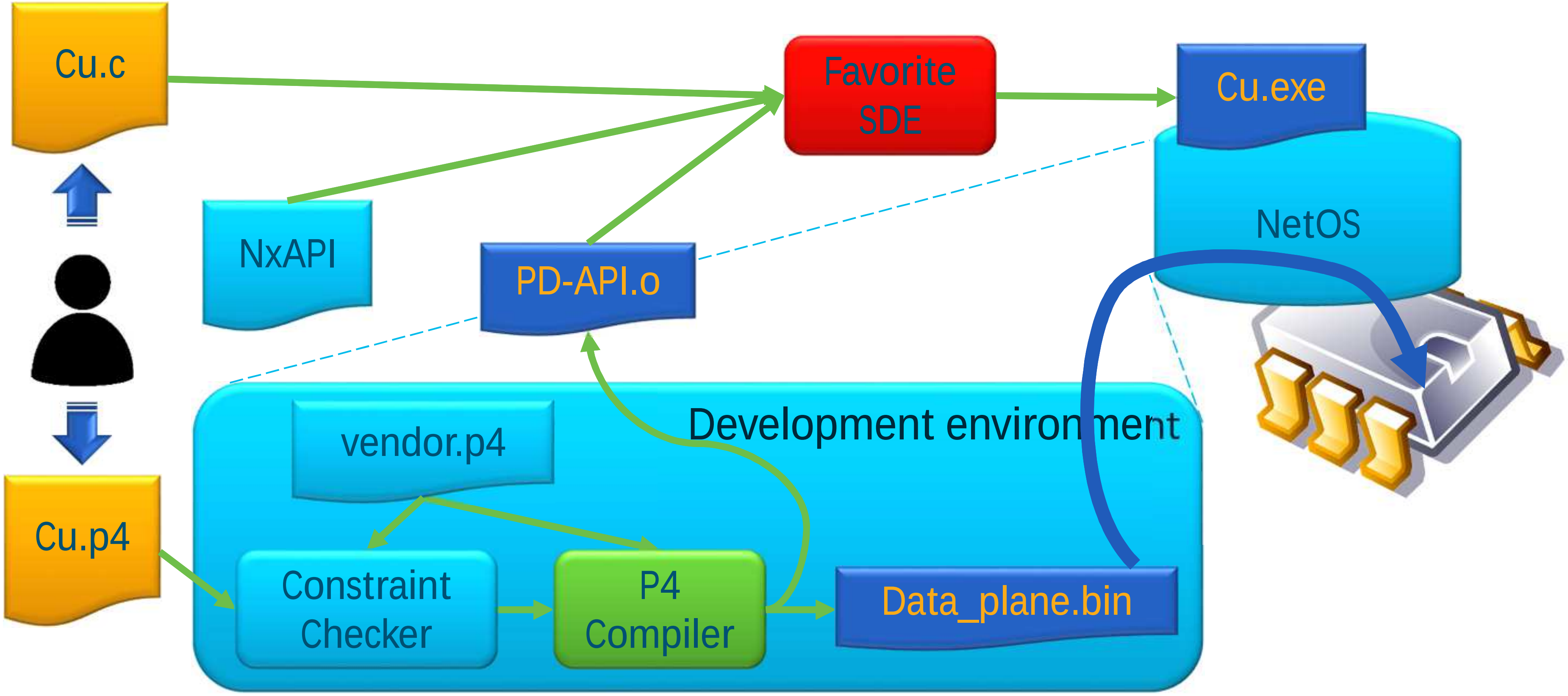
daPIPE build environment

Control pr

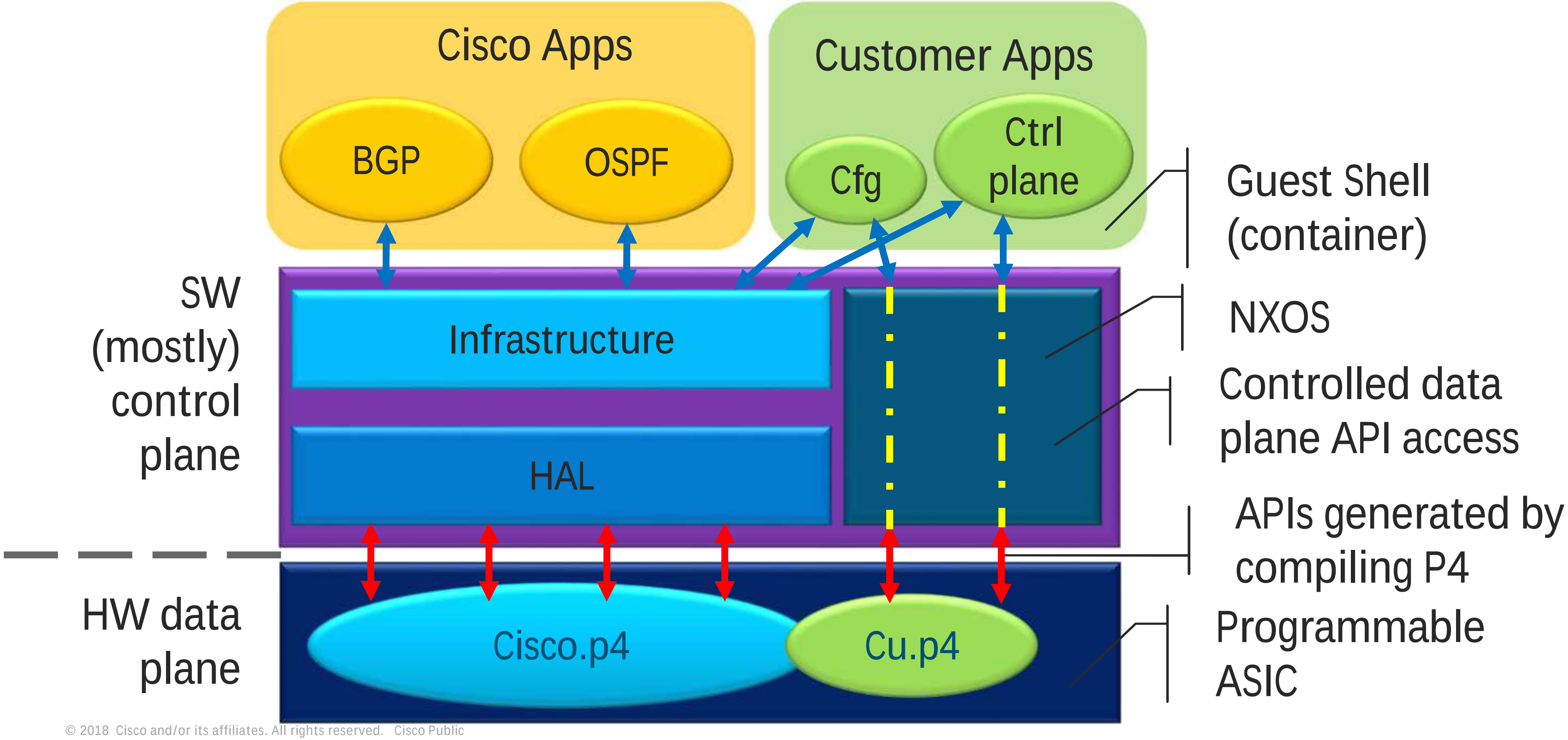
daPIPE Graphical User Interface

Nexus 3400

Customer Programming Workflow



Operating System Support


Cisco Apps
BGP
OSPF
Customer Apps
Cfg
Ctrl plane
Guest Shell (container)
SW (mostly) control plane
Infrastructure
HAL
NXOS
Controlled data plane API access
APIs generated by compiling P4
HW data plane
Cisco.p4
Cu.p4
Programmable ASIC