Cisco live!

What You Make Possible

TOMORROW
starts here.

cisco

Exploring the Engineering Behind the Making of a Switch

BRKARC-3466

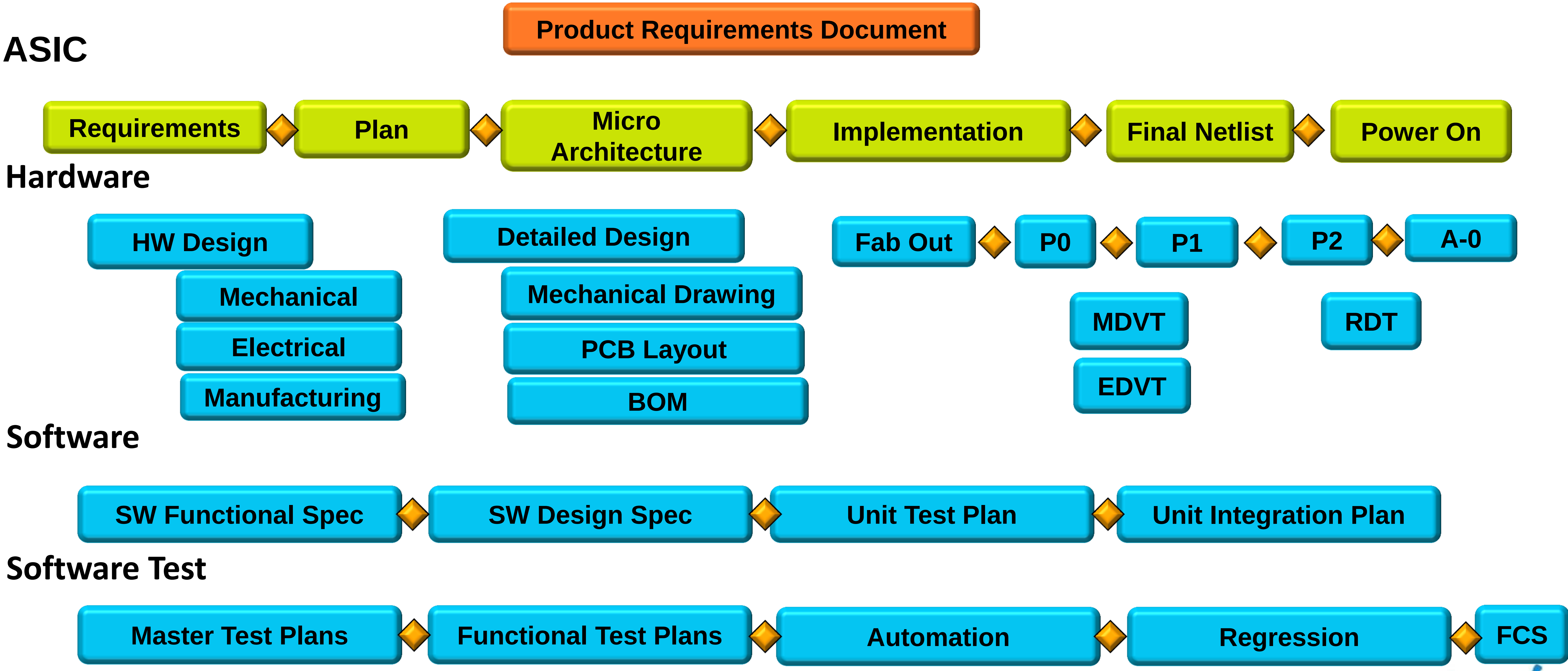
TOMORROW starts here.

Agenda

- Overview
- Concept
- System Design
- Mechanical / Physical Design
- Buffer Design
- Forwarding Design
- ASIC Engineering
- Hardware Engineering
- Software Engineering

Overview

Timeline



Product Requirements Document

ASIC

Requirements → Plan → Micro Architecture → Implementation → Final Netlist → Power On

Hardware

HW Design

Mechanical

Electrical

Manufacturing

Detailed Design

Mechanical Drawing

PCB Layout

BOM

Fab Out → P0 → P1 → P2 → A-0

MDVT

EDVT

RDT

Software

SW Functional Spec → SW Design Spec → Unit Test Plan → Unit Integration Plan

Software Test

Master Test Plans → Functional Test Plans → Automation → Regression → FCS

Nexus 7000 and F2/F2E Modules

Concept

Concept

What customer problem will the product solve?

- Vision
- Market
- Cost
- Time to Market
- Differentiation
- Innovation

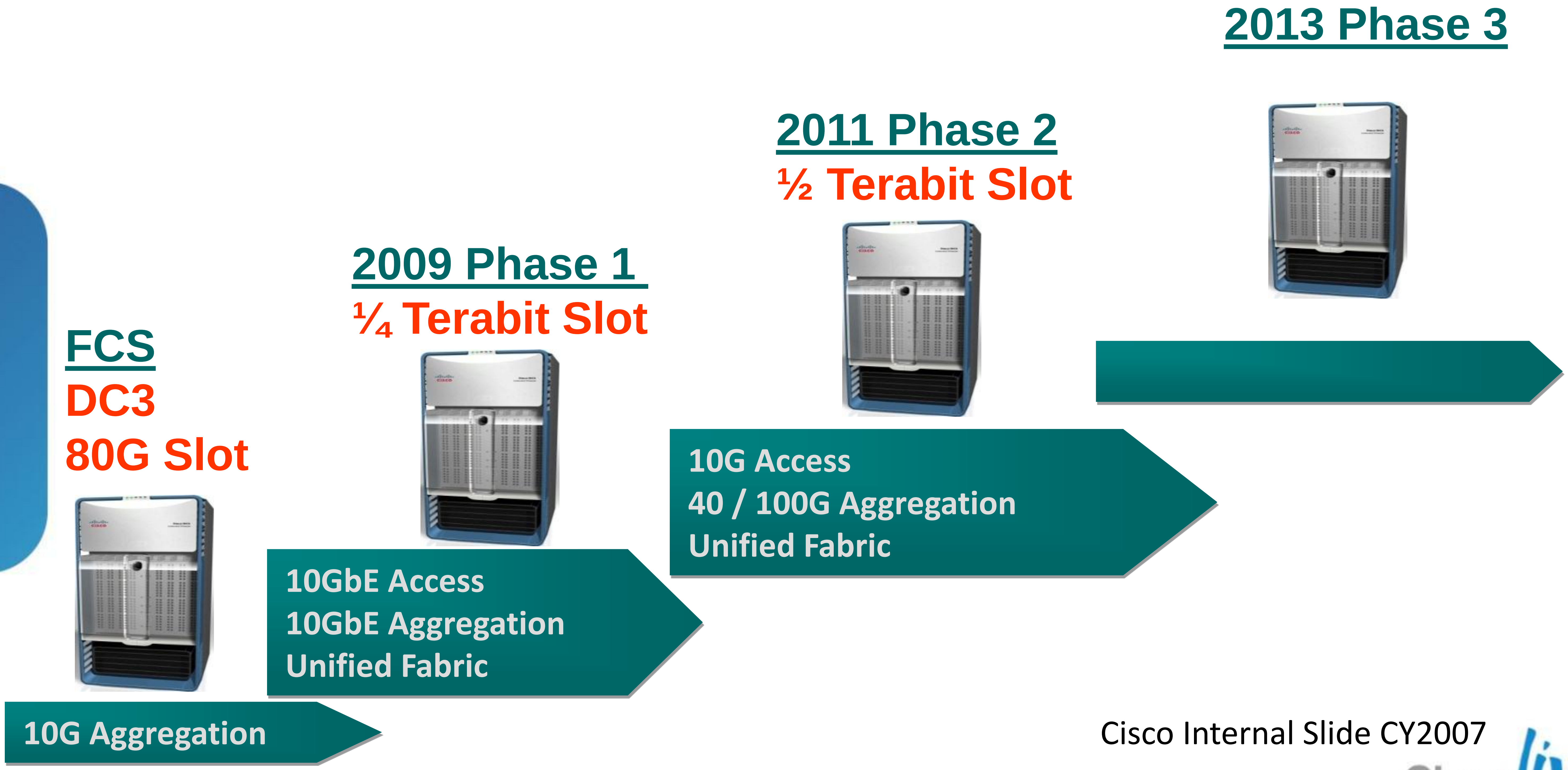
- Technology
- Life Cycle
- How Big?
- How many ports?
- Fixed vs Modular
- Backward Compatibility

Nexus 7000 Vision (circa 2007)

Cisco’s End-to-End Data Centre Switching platform; providing solutions for 10G, 40G, and 100G for Access, Aggregation, and Core.

Consolidate IP, Storage, and IPC networks onto a single Ethernet fabric and deliver innovative features and services that provide value to our customers.

DC Evolutionary Innovation



FCS

DC3

80G Slot

10G Aggregation

2009 Phase 1

¼ Terabit Slot

10GbE Access
10GbE Aggregation
Unified Fabric

2011 Phase 2

½ Terabit Slot

10G Access
40 / 100G Aggregation
Unified Fabric

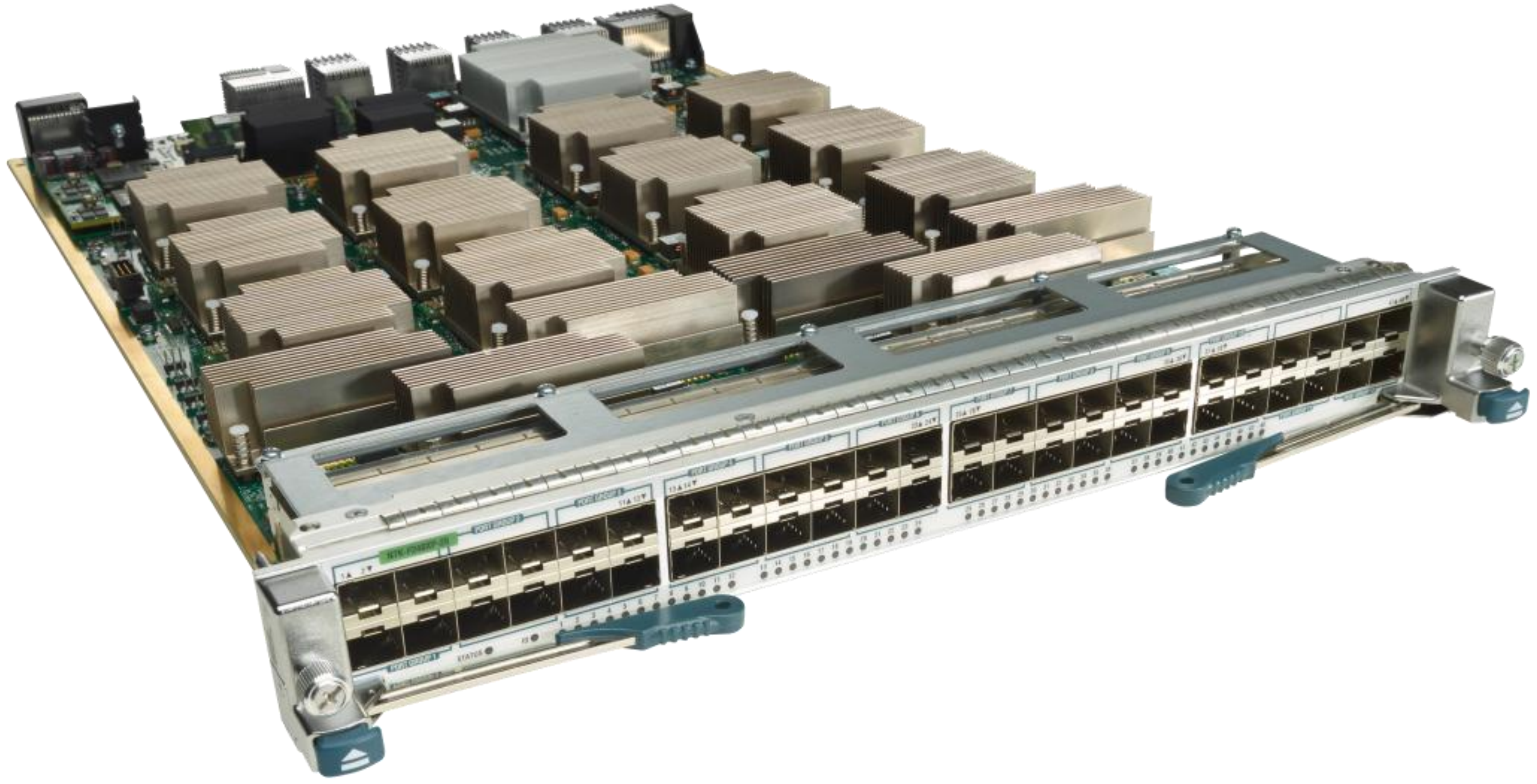
2013 Phase 3

Cisco Internal Slide CY2007

F2 Series

High Level Goals

- 48 Ports 1/10G
  Line Rate 64 Bytes
- Low Latency
- L2MP, TRILL, FEX, FCoE, L3 Forwarding
- Optimise for Data Centre
- IPv4 & IPv6
  Equal Performance
- Cost target



System Design

Many Factors to Weigh

Applicable to any Switch / Router Design

- Standards requirements
- Market requirements
- Designability
- Silicon technology
- Processor technology

- Manufacturability
- Time to market
- Flexibility
- Budget
- Modular / Fixed

Many Factors to Weigh

Baseline Data Centre Switch Requirements

Data Plane

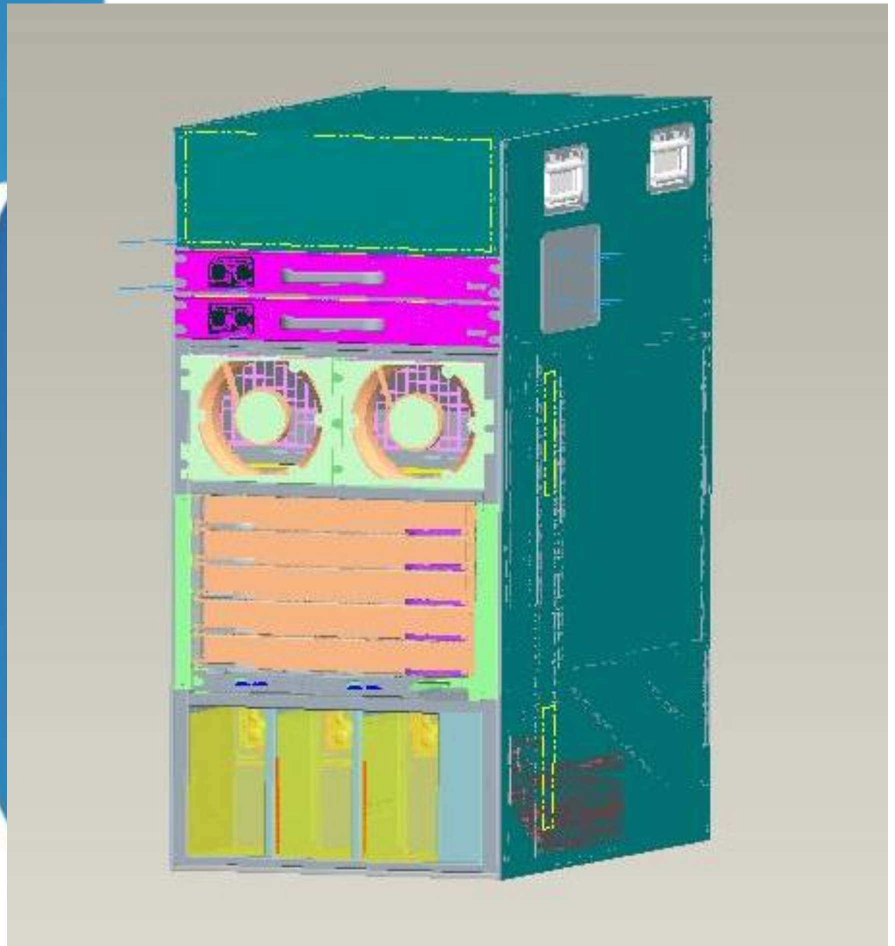
- Buffering
- No packet drop
- Throughput
- Port count
- Modular
- No single point of failure
- In-order delivery
- Future protocol compatibility

Control plane

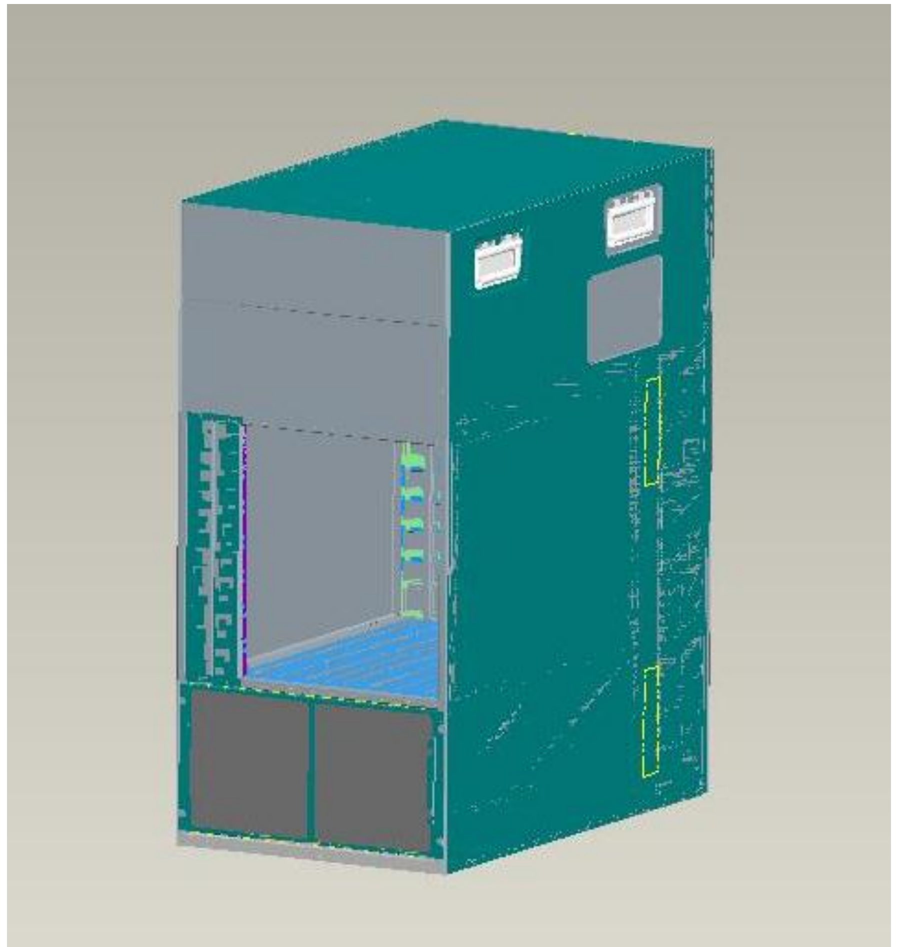
- Modular
- Restartable (including active-active state handling)
- Non-disruptive code load & activation
- No single point of failure
- Scaleable
- Unit Testable
- Future protocol compatibility

Mechanical / Physical Design

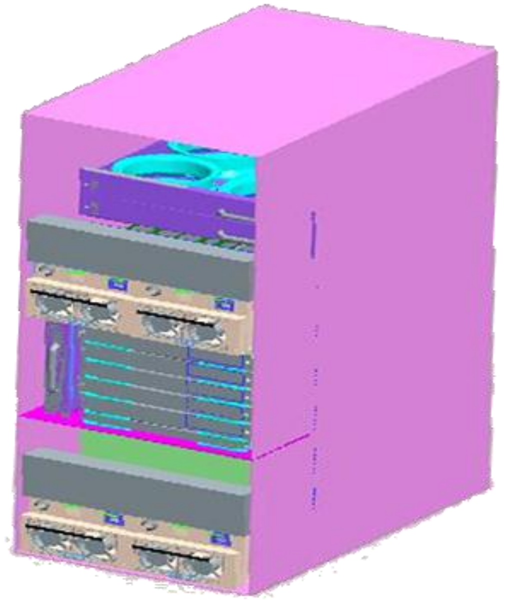
Mechanical Design



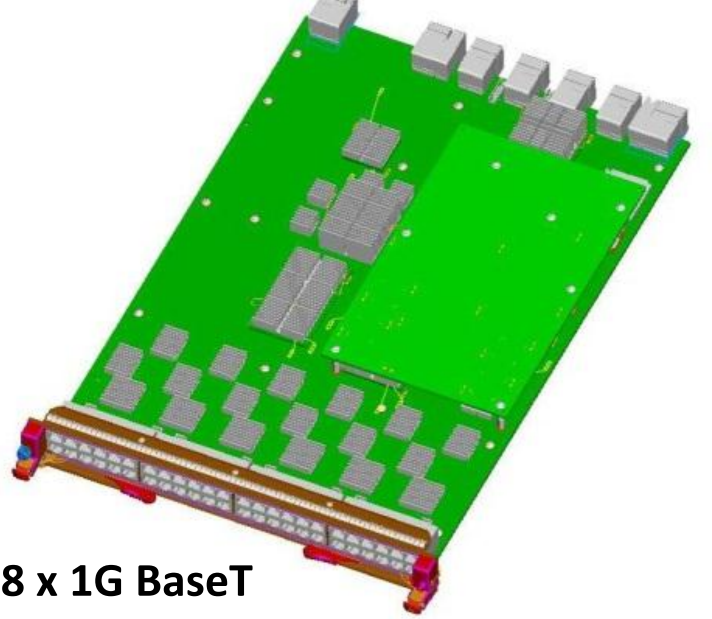
Nexus 7010 Rear



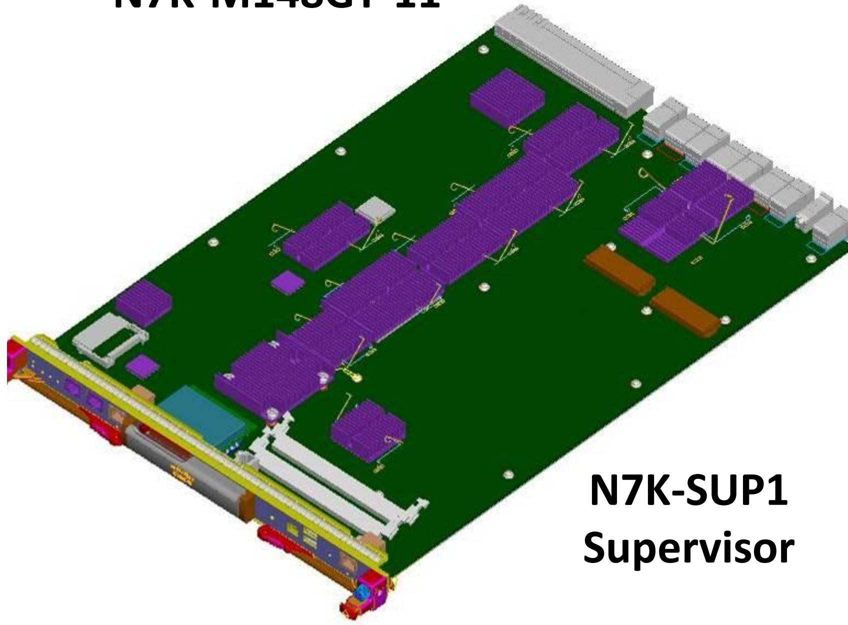
Nexus 7010 Front



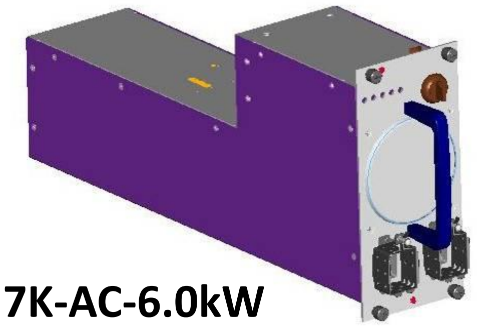
Nexus 7010 Rear



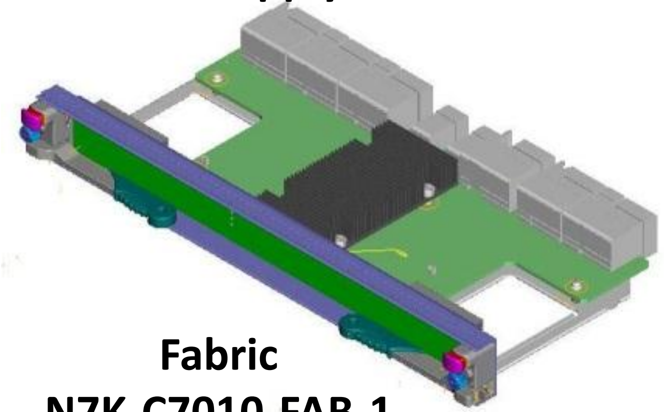
48 x 1G BaseT
N7K-M148GT-11



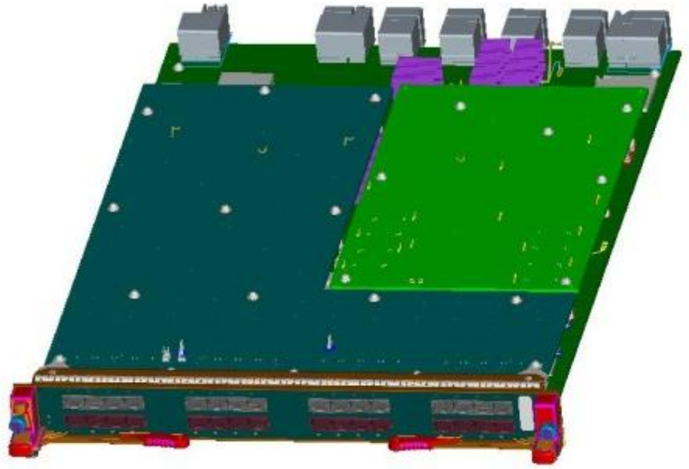
N7K-SUP1
Supervisor



N7K-AC-6.0kW
Power Supply



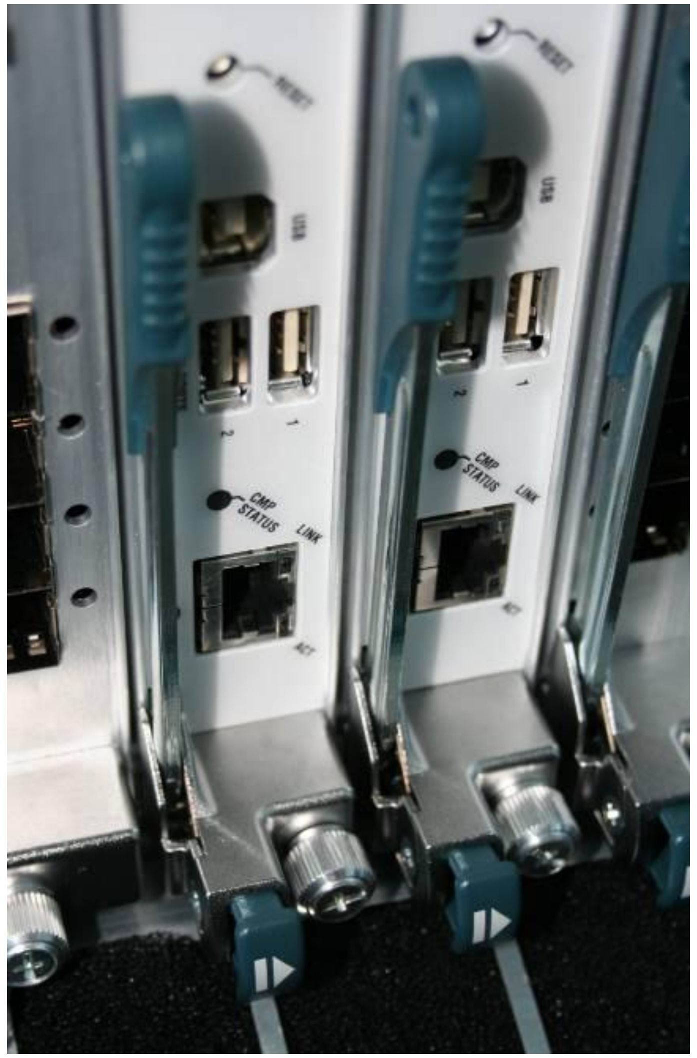
Fabric
N7K-C7010-FAB-1



32 x 10G SFP+
N7K-M132XP-12

Industrial Design / Usability

cisco
Cisco Systems
Cisco Systems DC3 and Cable Management Industrial Design Program Phase 2
Studio RED



Ejectors

Industrial Design / Usability

Buffer Design

Memory Technology

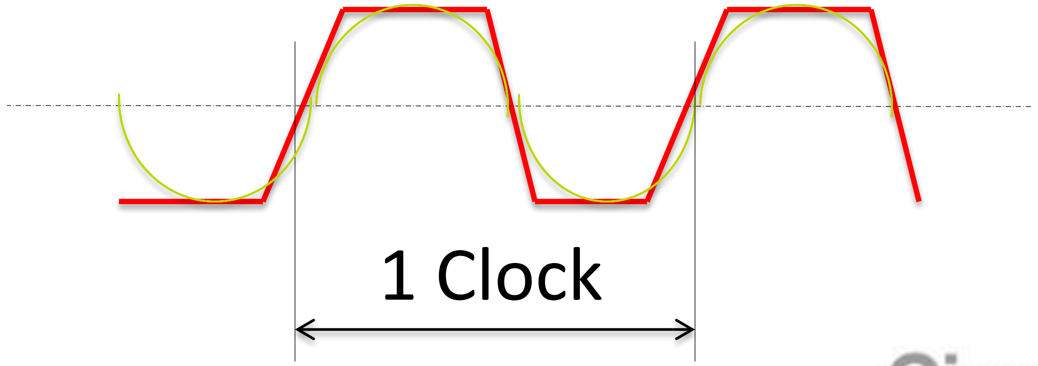
DRAM: Dynamic Random Access Memory
SDRAM: Synchronous Dynamic Random Access
eDRAM: embed Dynamic Random Access Memory

DDR3 Latency ~10ns
1 Transistor + 1 Capacitor
Requires Refresh

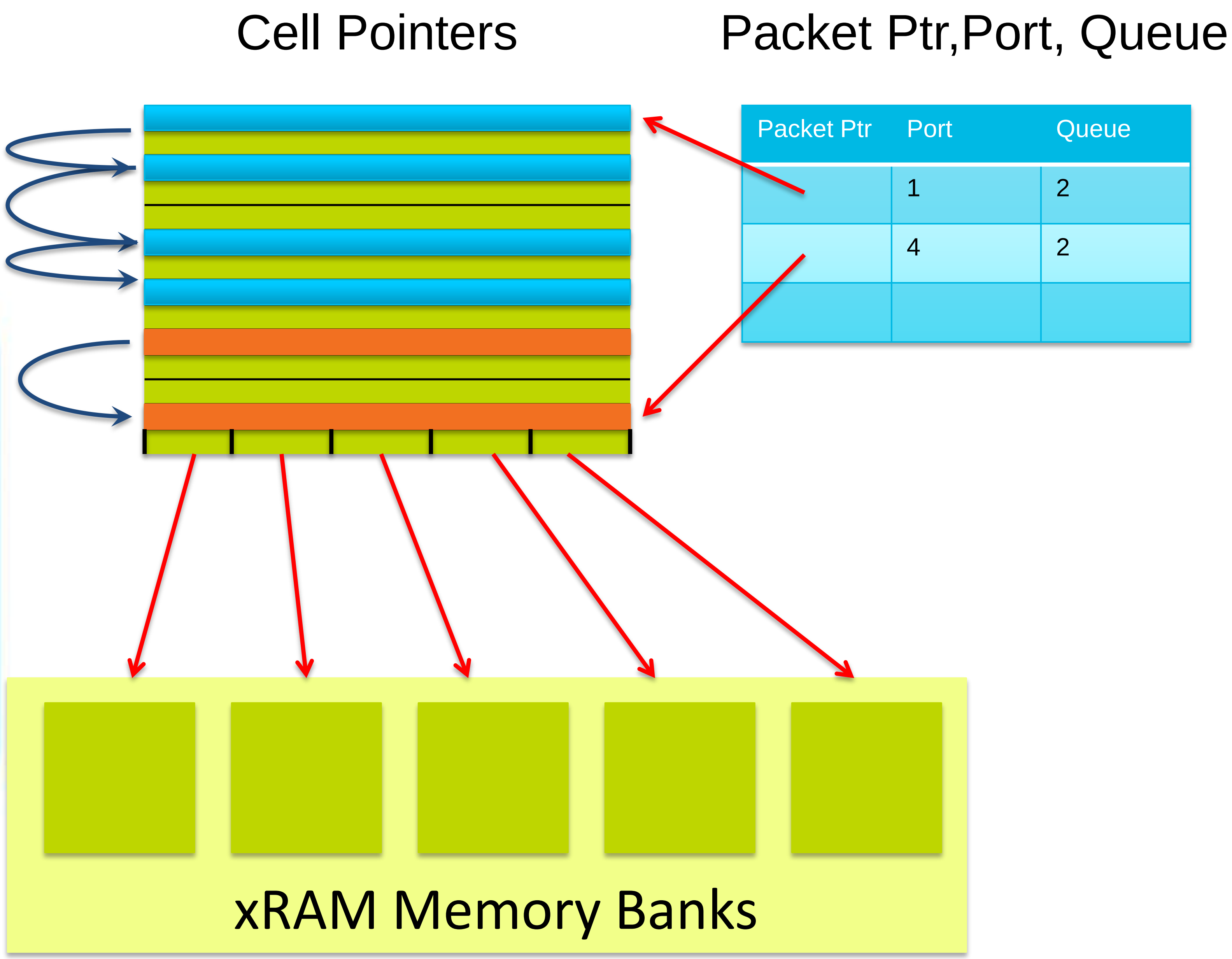
SRAM: Static Random Access Memory
SSRAM: Synchronous Static Random Access

SRAM Latency 1 cycle
6 Transistors

DDR: Double Data Rate – Transfer on Rising and Falling Edges of Clock
QDR: Quad Data Rate – Transfers on Rising and Falling and 2 intermediate points between them



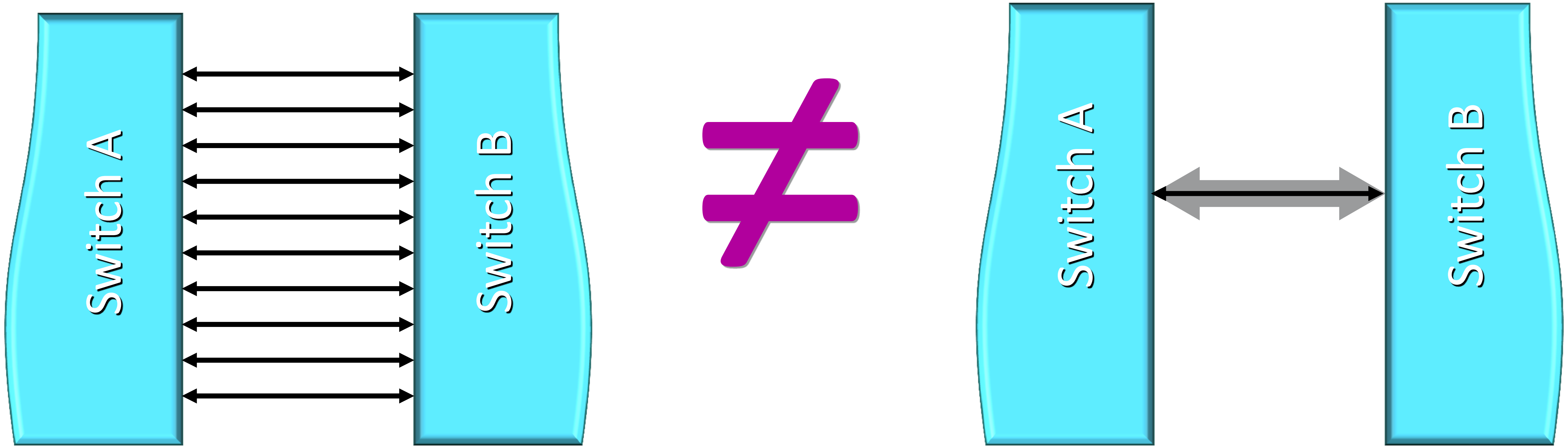
Packet Buffer Design



Packet Ptr	Port	Queue
	1	2
	4	2

- Fixed
  - Memory segmented into a fixed cell size, like 128, 384, 512 bytes
  - If packet is smaller than cell size, then the left over space in page is unused
  - Easier to map to different memory banks
- Packed
  - Packets are placed head to tail into the packet memory
  - More efficient utilisation of the memory
  - More complex free space management and bank mapping

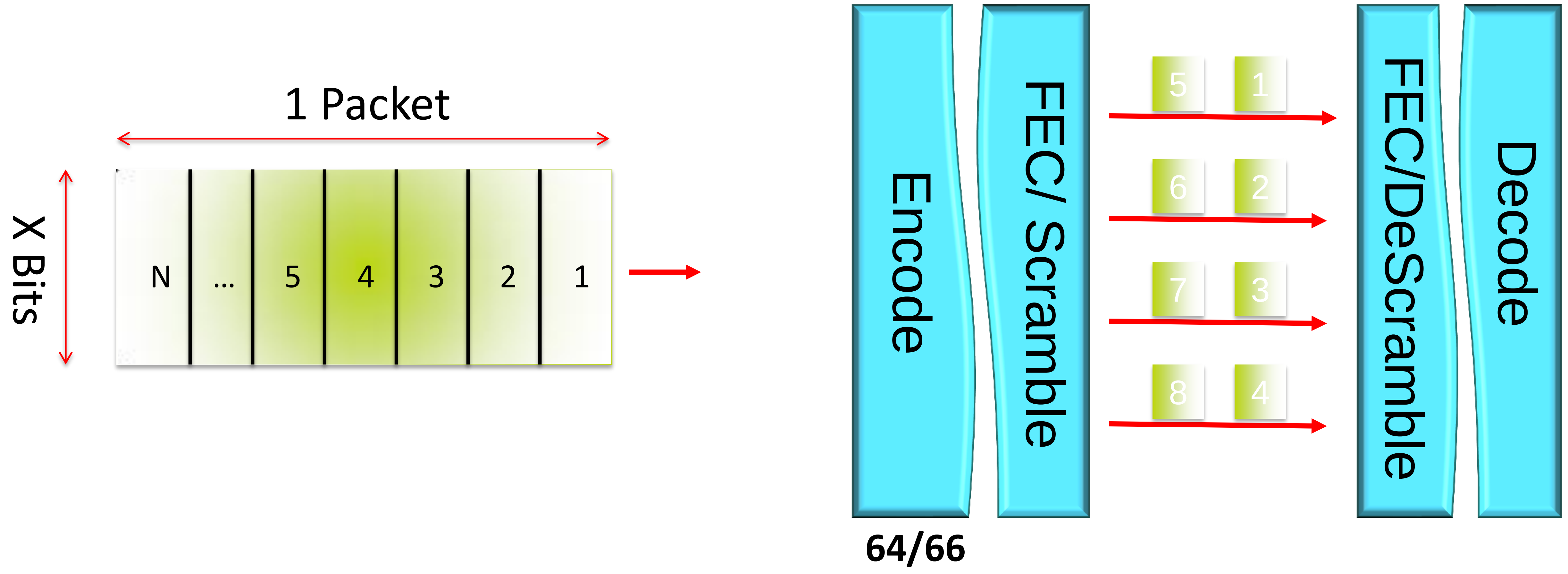
One plus one does not equal Two
Flow Balancing



10 Links @ 1Gbps Each
Bandwidth = 10Gbps
Flow Bandwidth = 1Gbps
Serialisation Delay = 20uS

1 Link @ 10Gbps Each
Bandwidth = 10Gbps
Flow Bandwidth = 10Gbps
Serialisation Delay = 2uS

One Plus One Equal Two
Word Spraying



Creates a link N x link speed, with small increase in Latency due to supporting difference in delays between physical lanes

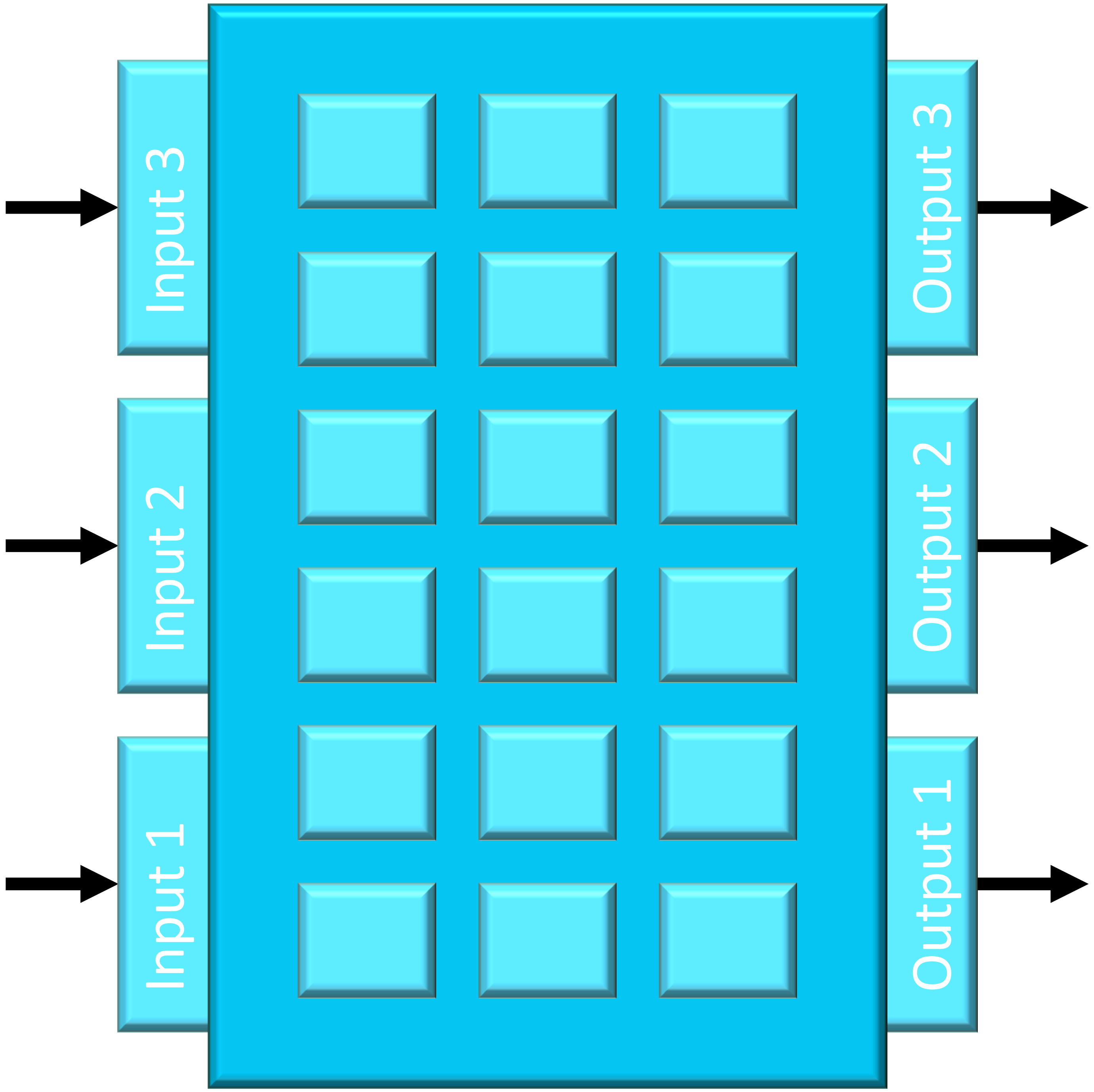
Example 40/100G Ethernet, 180ns maximum skew between lanes

Scrambing / Encoding

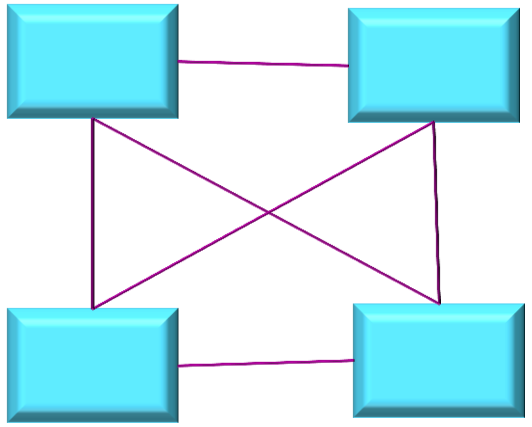
	Serdes (Gbps)	Encoding	Bandwdith after Encoding (Gbps)
PCI express v1	2.525	8b/10b	2.02
PCI express v2	5G	8b/10b	4
PCI express v3	7.99	128b/130b	7.867
10G Ethernet XAUI	4 x 3.126G	8b/10b	10
10G Ethernet XFI	10.3125	64b/66b	10
40G Ethernet 4x XFI	4x 10.3125	64b/66b	40
100G Ethernet	10 x 10.3125	64b/66b	100
100G Ethernet 4x 25G	4x 25.78125G	64b/66b	100
2G FC	2.125	8b/10b	1.7
4G FC	4.250	8b/10b	3.4
8G FC	8.5	8b/10b	6.8
16G FC	14.025	64b/66b	13.6

Single ASIC

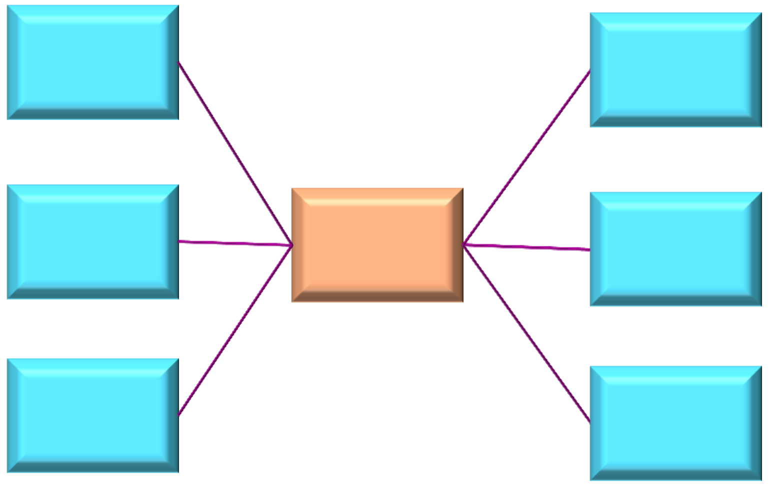
- Scalability limited by memory bandwidth/size
- Typically optimised for fixed configuration
- Cost effective with small port counts
- Often used as building block



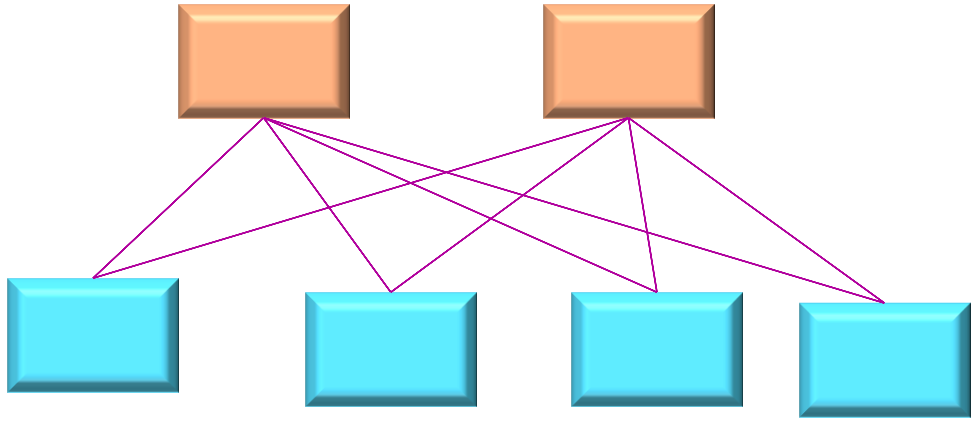
Switch Architecture



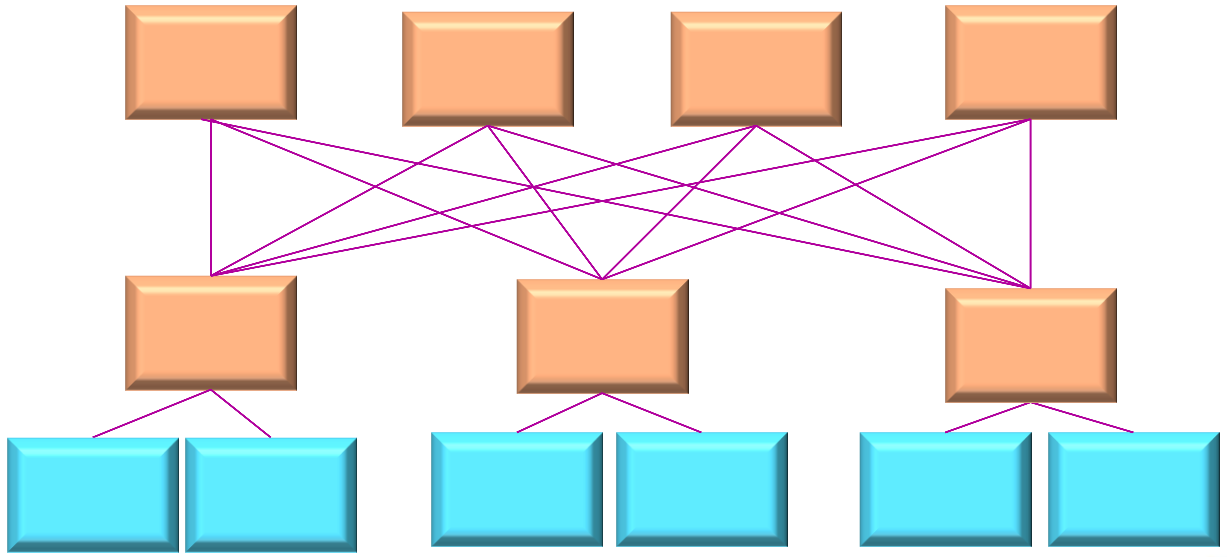
Mesh



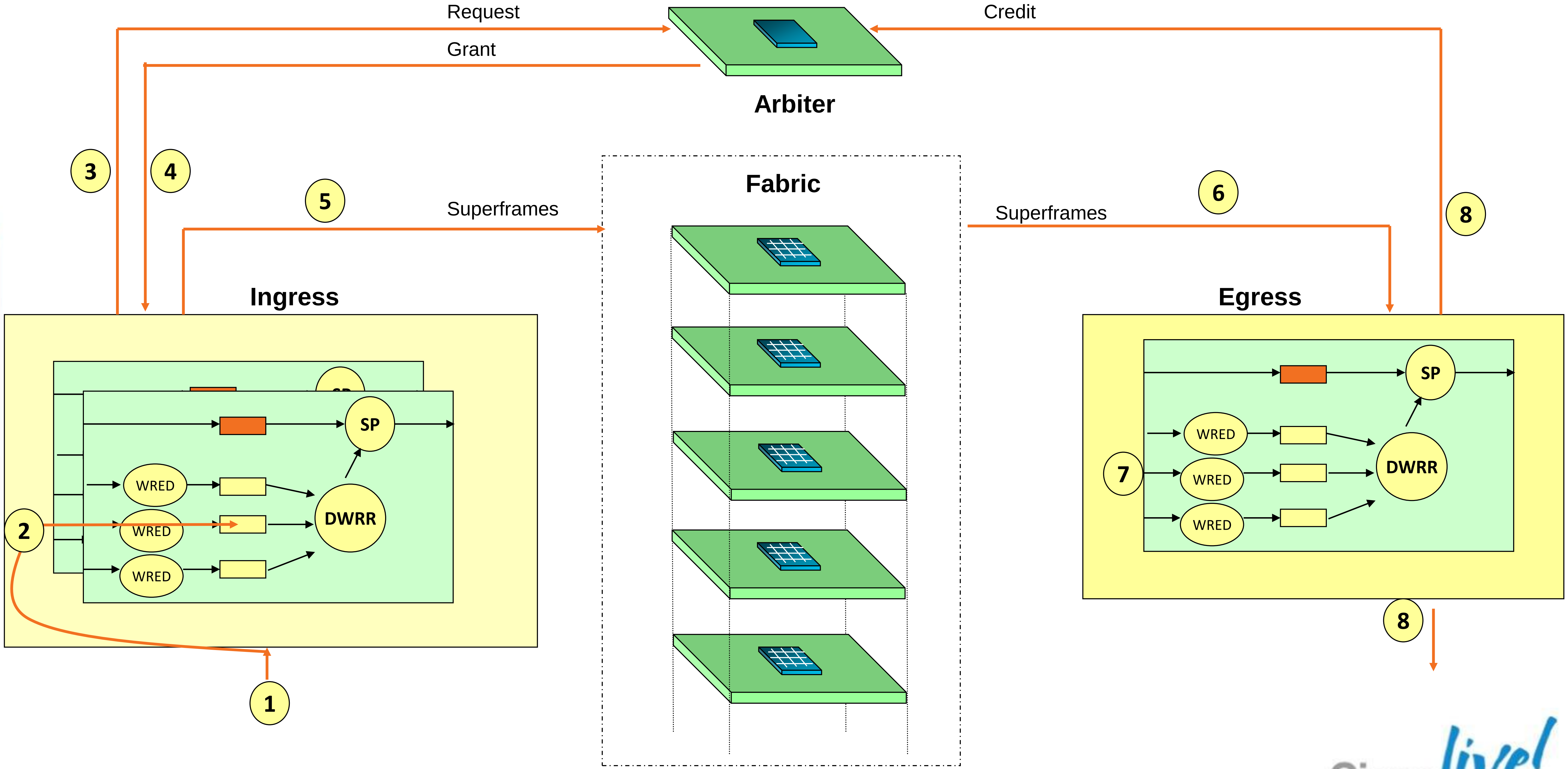
Crossbar



Clos / Fat Tree

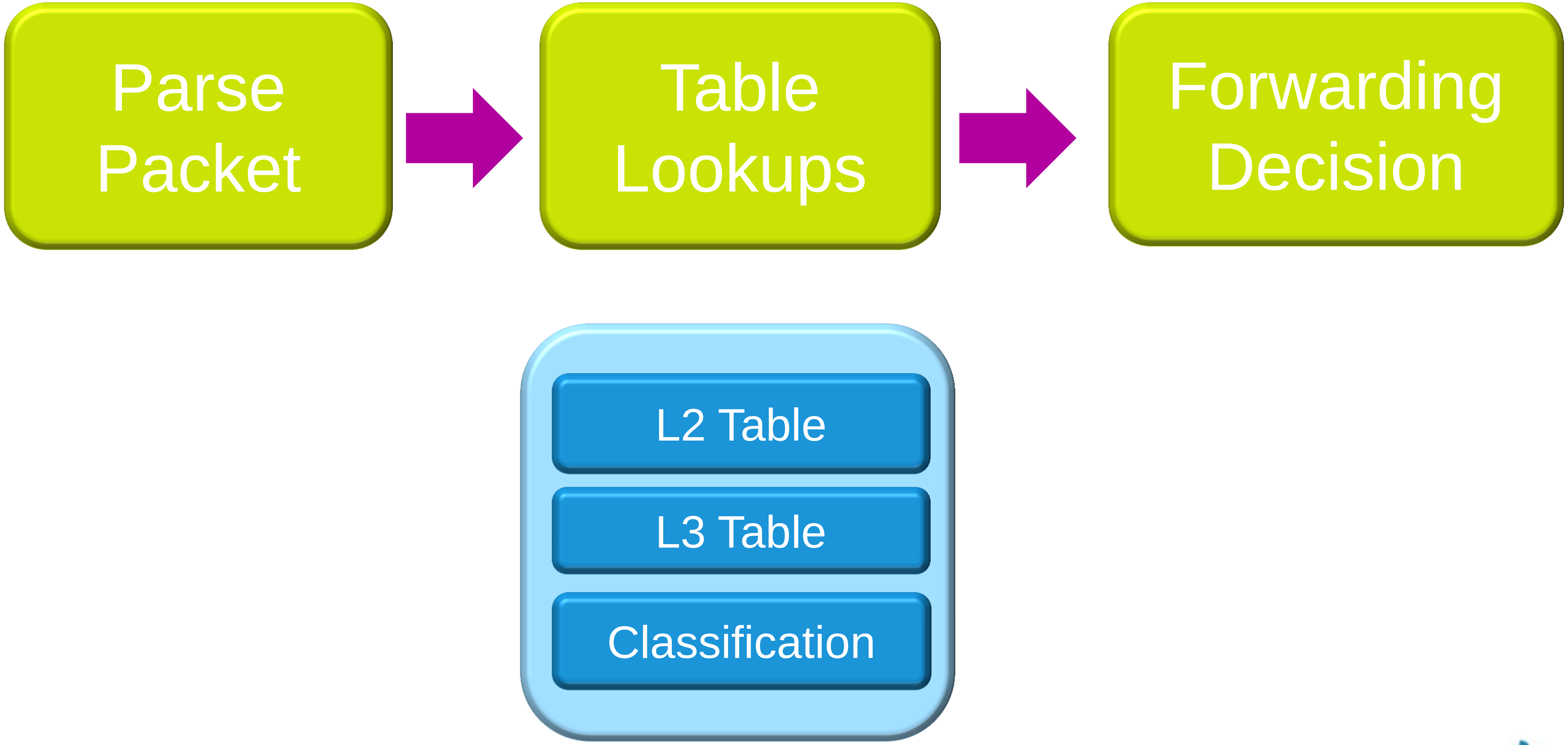


Complete System – Pull Fabric

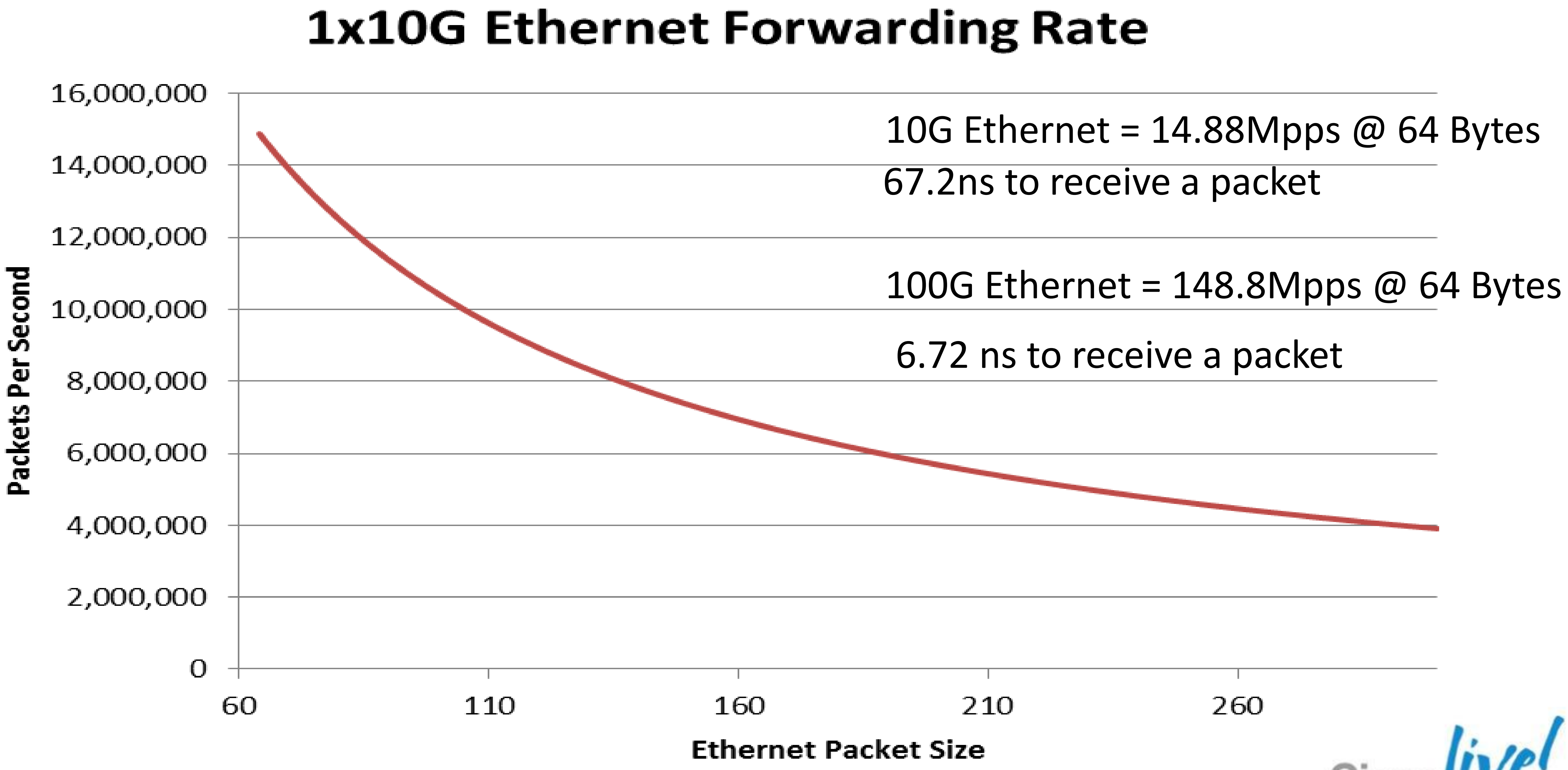


Forwarding Design

High Level View of Forwarding



10G Ethernet Forwarding Rate



10G Ethernet = 14.88Mpps @ 64 Bytes

67.2ns to receive a packet

100G Ethernet = 148.8Mpps @ 64 Bytes

6.72 ns to receive a packet

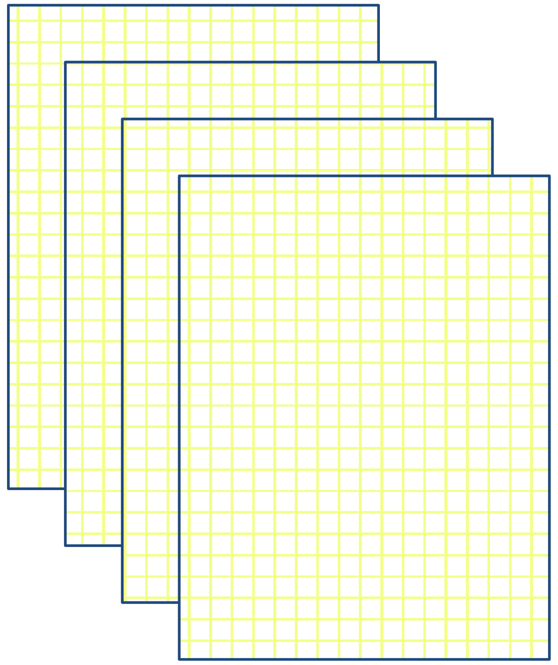
Table Lookups
CAMs, HASH Tables and *Tries

Input Key

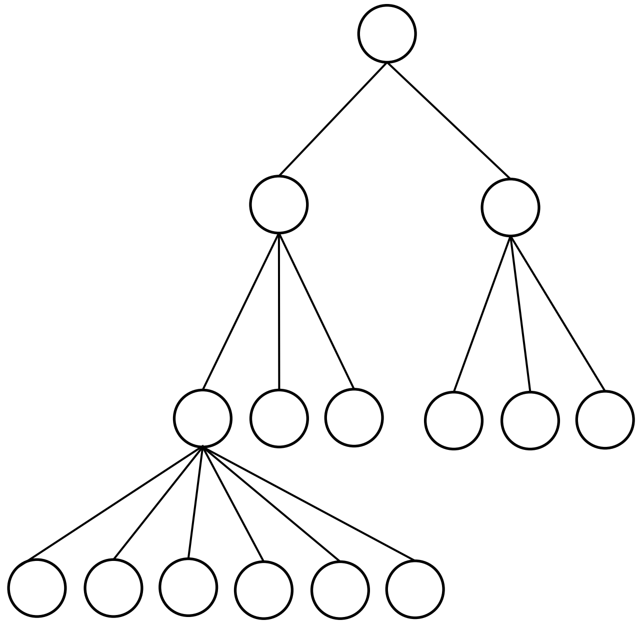
TCAM

1	01001010
2	010010XX
3	01001XX0
4	01001XXX

Hash Table



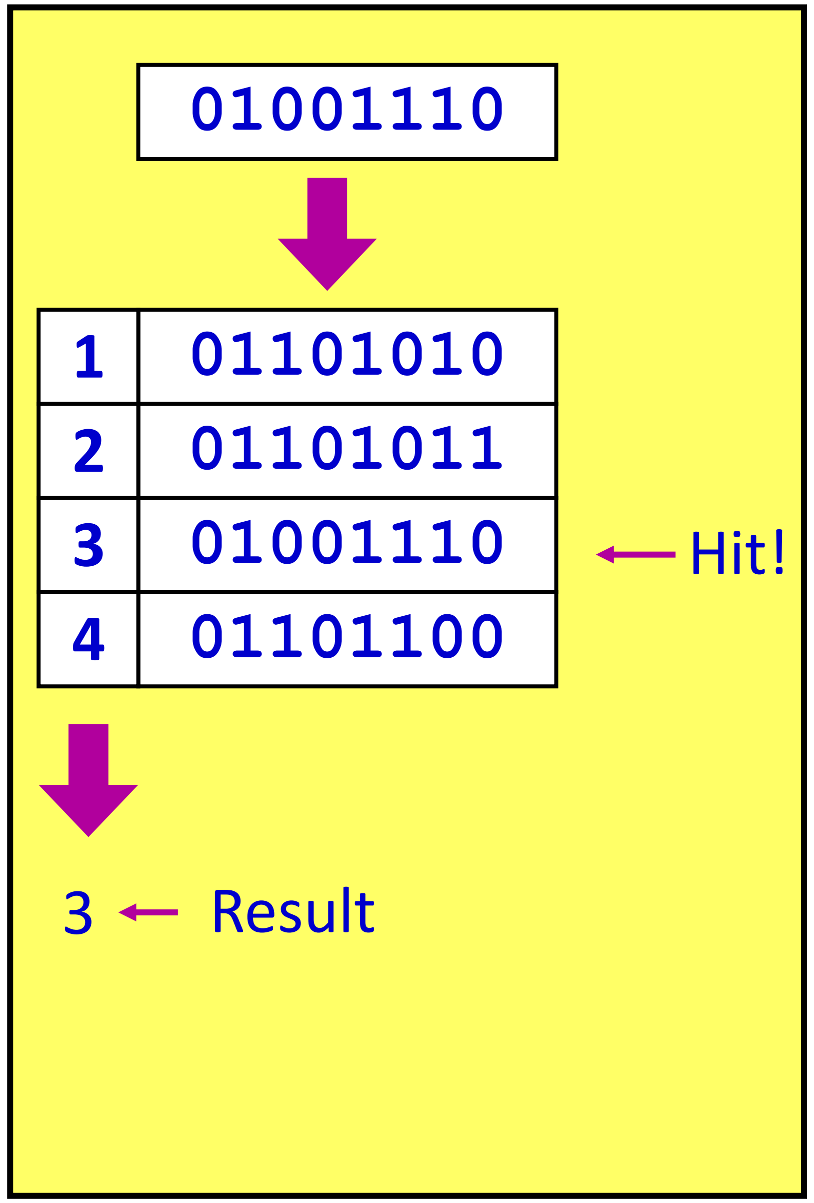
Trie



Result

CAMs

Content Addressable Memory



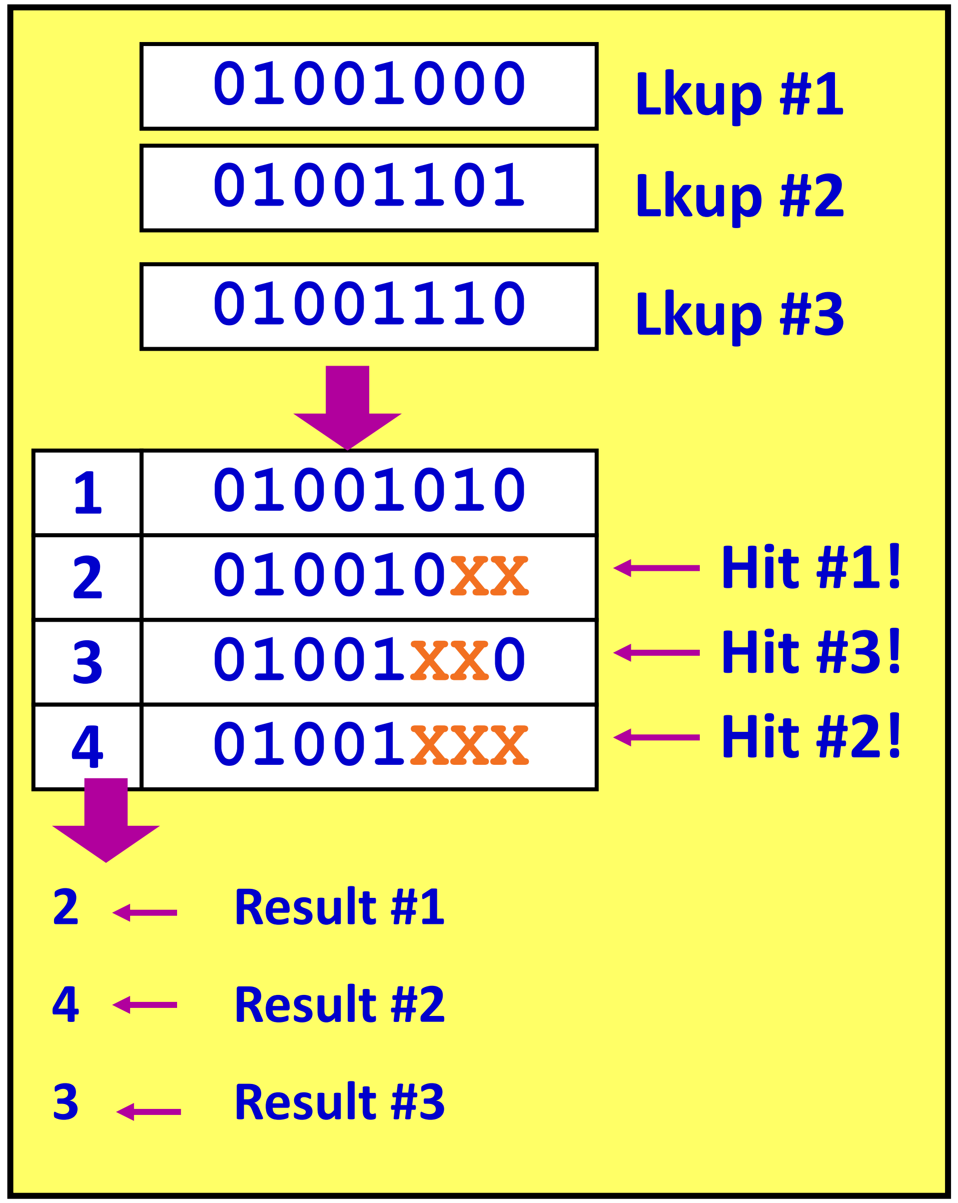
01001110

1	01101010
2	01101011
3	01001110
4	01101100

← Hit!

3 ← Result

Ternary Content Addressable Memory



01001000 Lkup #1

01001101 Lkup #2

01001110 Lkup #3

1	01001010	
2	010010XX	← Hit #1!
3	01001XX0	← Hit #3!
4	01001XXX	← Hit #2!

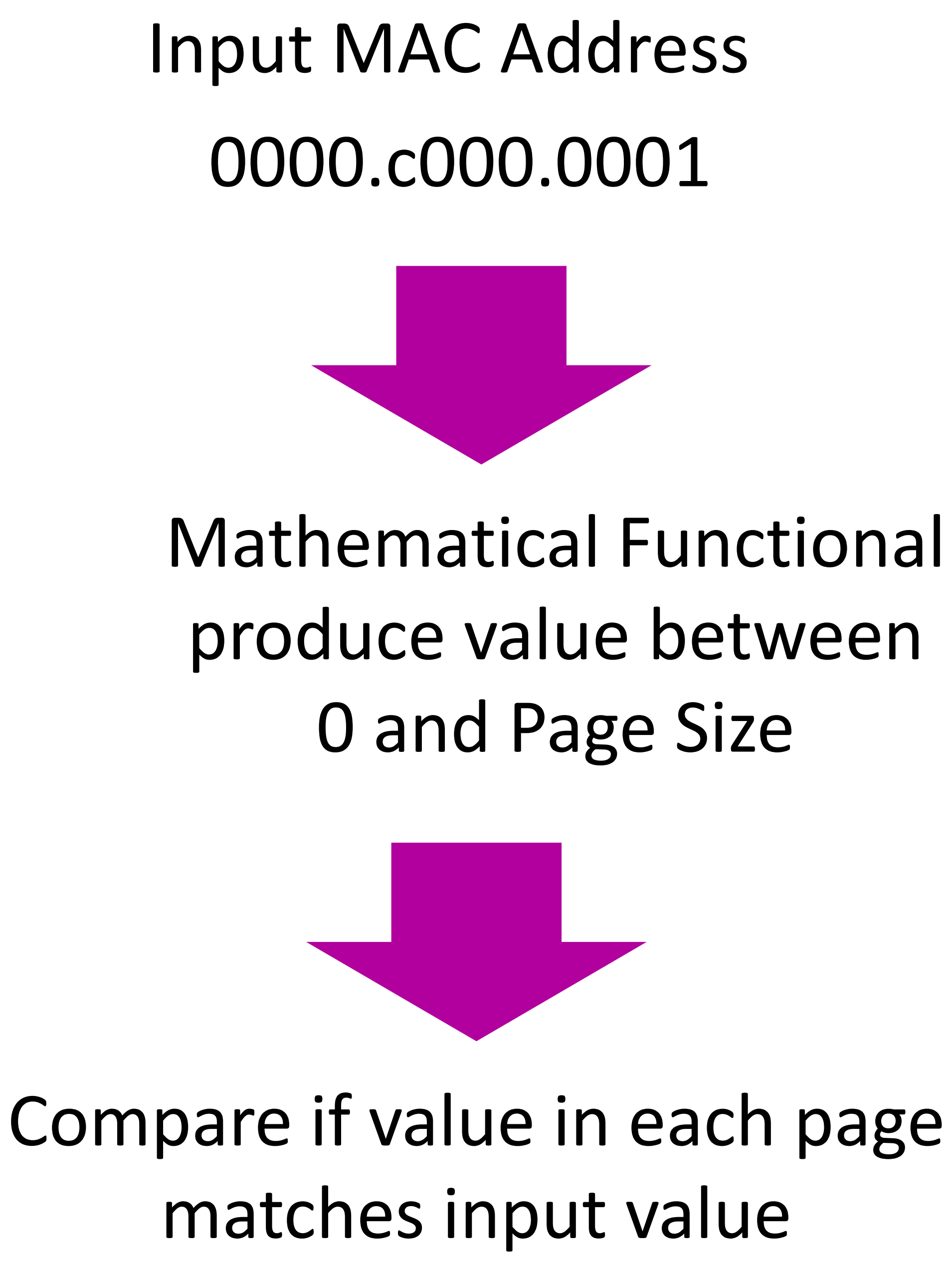
2 ← Result #1

4 ← Result #2

3 ← Result #3

Storing 1 bit in TCAM takes 10-12 transistors

Hash Tables

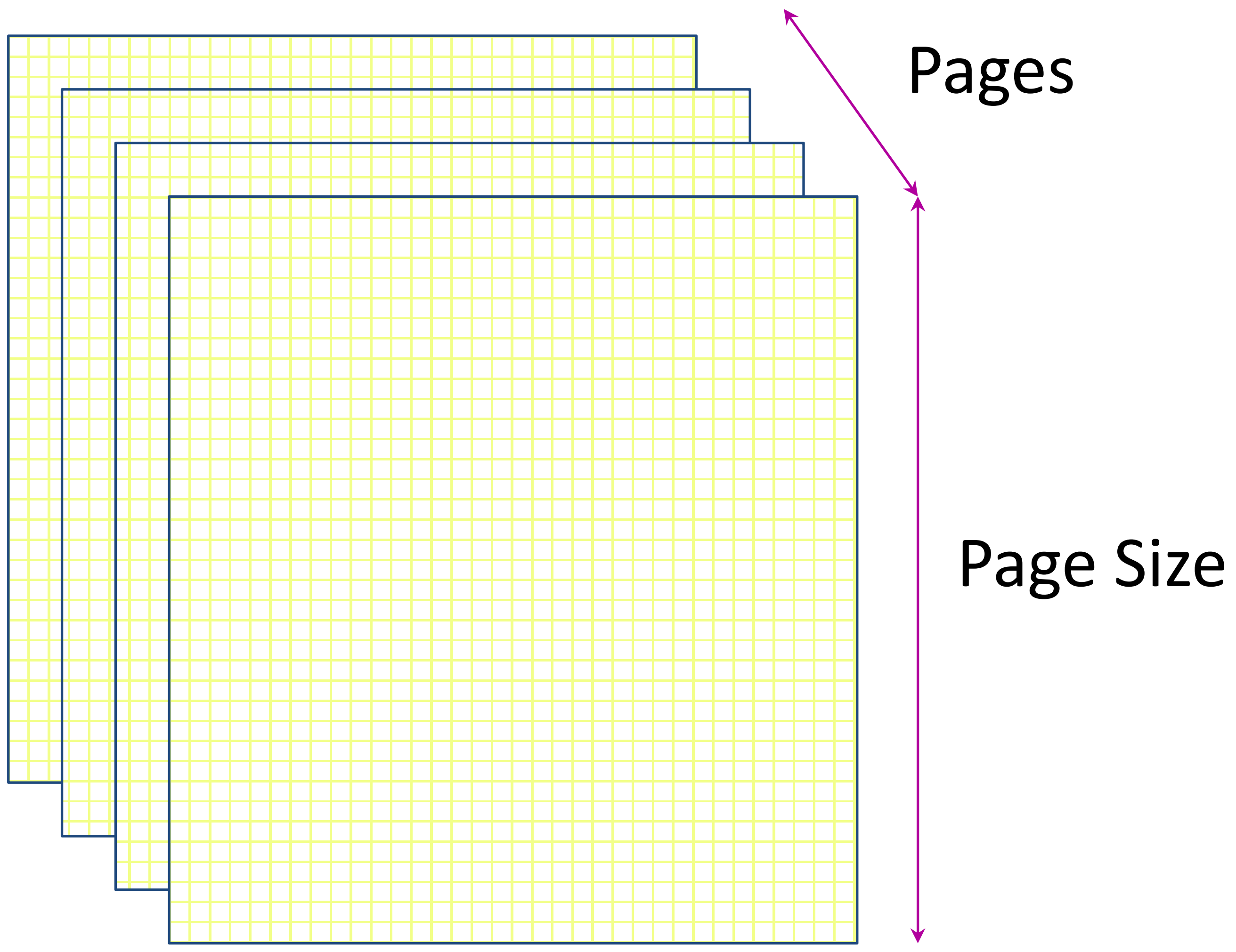


Input MAC Address

0000.c000.0001

Mathematical Functional produce value between 0 and Page Size

Compare if value in each page matches input value

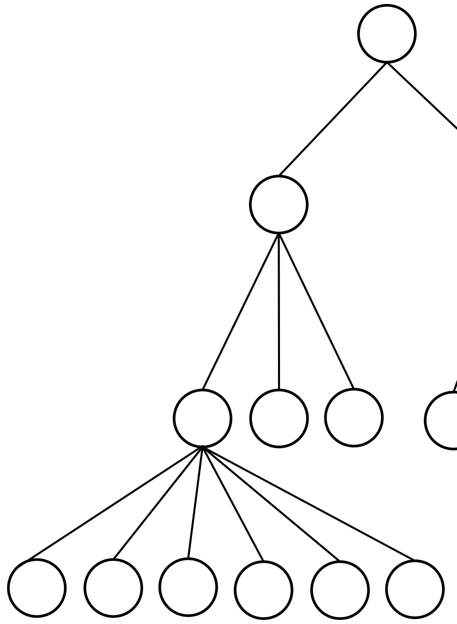


1 bit in SRAM takes 6 transistors

1 bit in DRAM takes 1 transistor

Tries

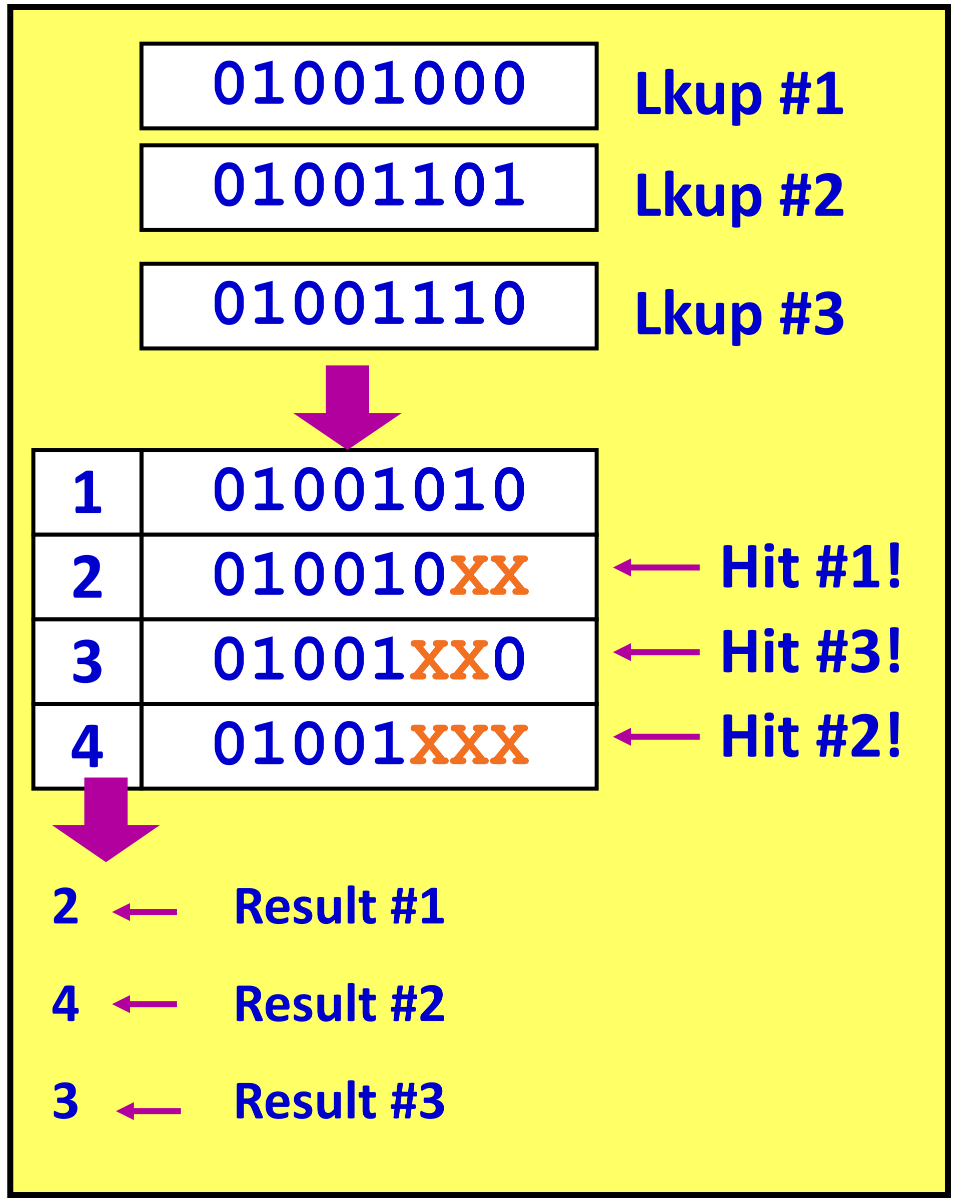
- Many different *tries
  - – Bitwise Trie
  - – Balanced Trie
  - – Patricia Trie
  - – Fixed or Variable Stride Tries
- Store information in each leaf or pointer to table with information in it



Algorithmic TCAMs

- From a software point of view looks like TCAM
  - May be all algorithmic or combination of TCAM and algorithmic
  - Software Driver takes TCAM representation and compiles the table to optimally utilise the underlying device
- Why? Algorithmic approaches allow the tables to scale with less than linear power increase

High Level View
Algorithmic TCAM

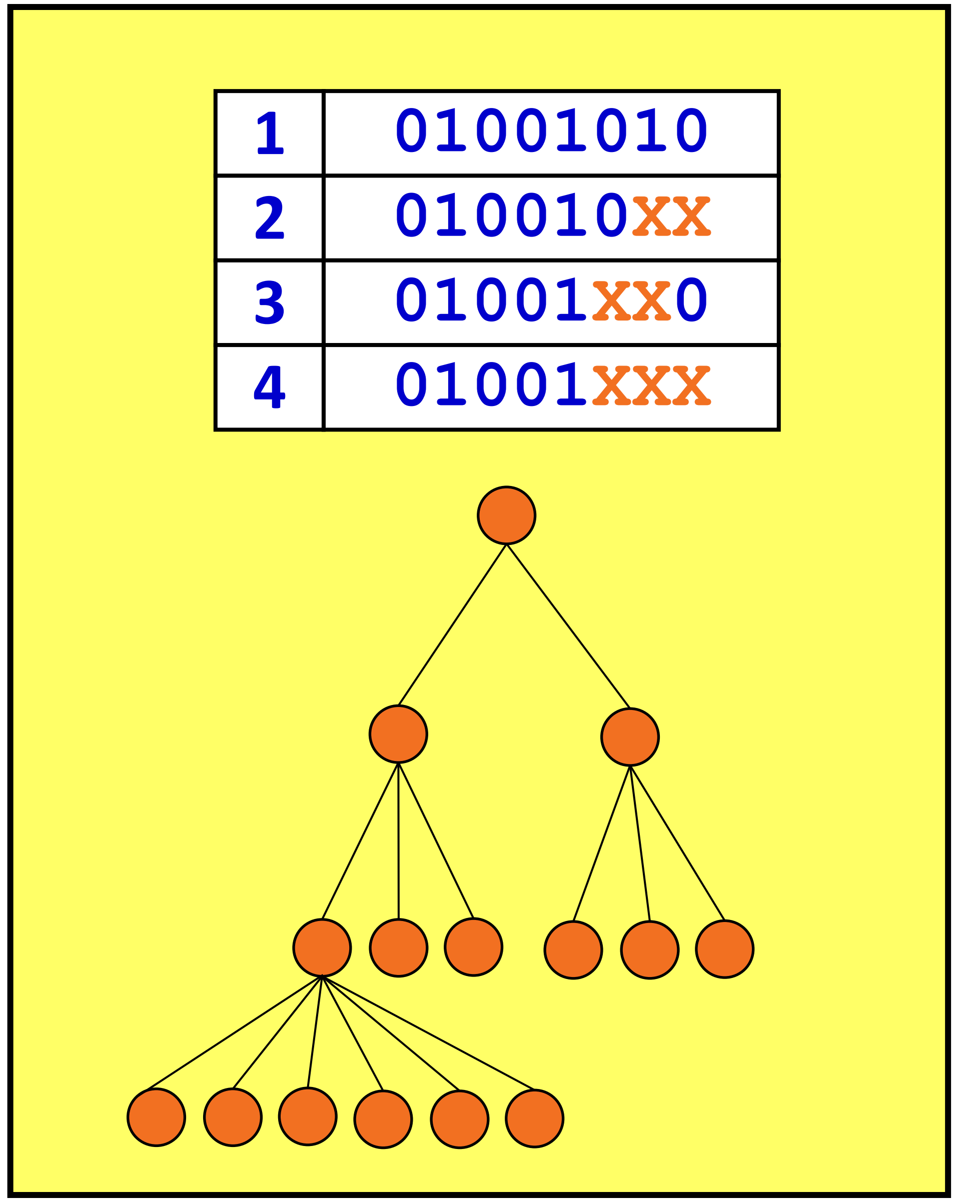


01001000	Lkup #1
01001101	Lkup #2
01001110	Lkup #3

1	01001010	
2	010010XX	Hit #1!
3	01001XX0	Hit #3!
4	01001XXX	Hit #2!

2	Result #1
4	Result #2
3	Result #3

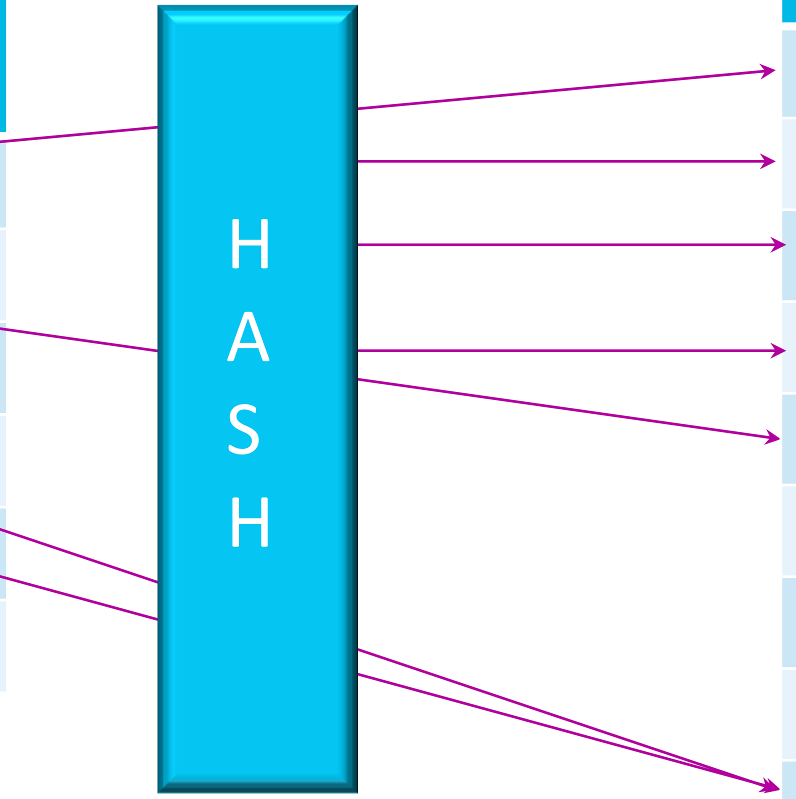
Implementation
Algorithmic TCAM



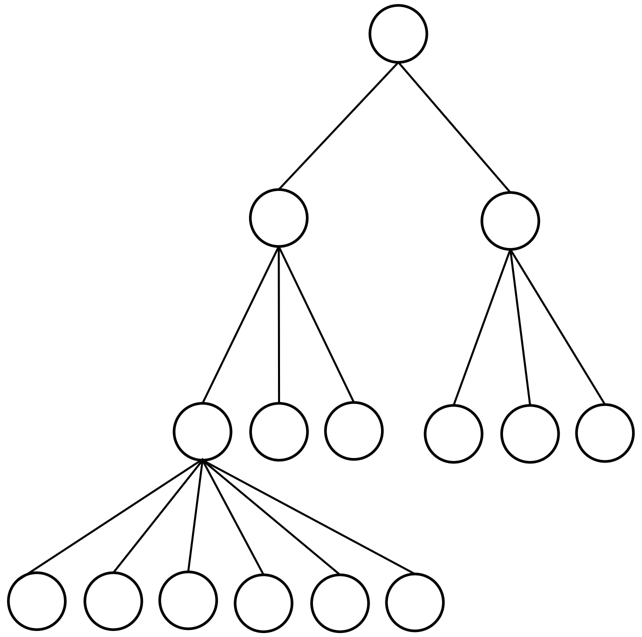
1	01001010
2	010010XX
3	01001XX0
4	01001XXX

L3 Table: Design 1

IPv4 Unicast FIB VRF / Prefix / Mask / Paths / Offset
1 / 10.1.2.0 / 24 / 4 / 1
1 / 10.1.3.0 / 24 / 1 / 5
3 / 10.1.2.0 / 24 / 2 / 9
3 / 10.1.3.0 / 24 / 2 / 9

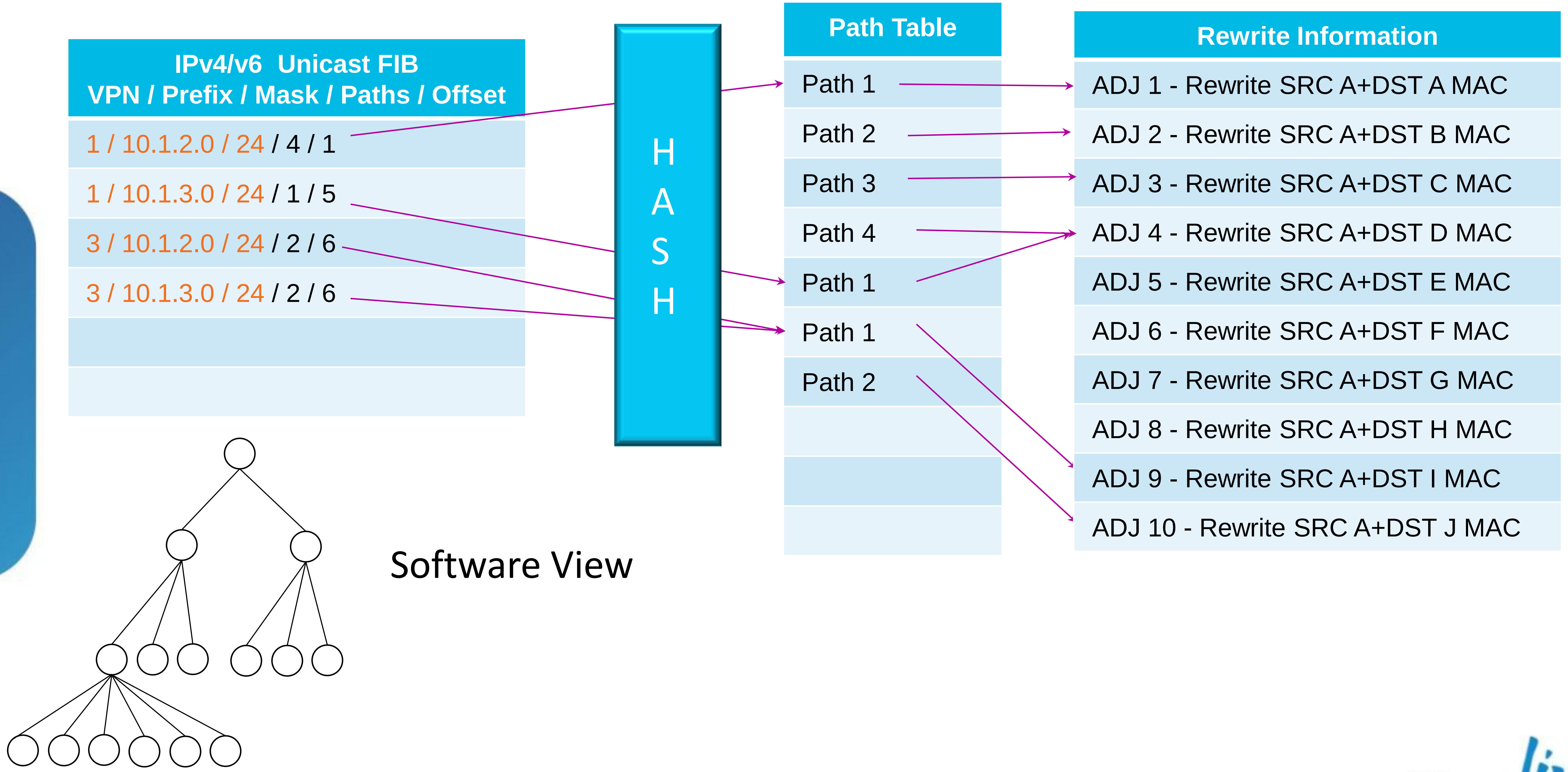


Rewrite Information
ADJ 1 - Rewrite SRC A+DST A MAC
ADJ 2 - Rewrite SRC A+DST B MAC
ADJ 3 - Rewrite SRC A+DST C MAC
ADJ 4 - Rewrite SRC A+DST D MAC
ADJ 5 - Rewrite SRC A+DST D MAC
ADJ 6 - Rewrite SRC A+DST F MAC
ADJ 7 - Rewrite SRC A+DST G MAC
ADJ 8 - Rewrite SRC A+DST H MAC
ADJ 9 - Rewrite SRC A+DST I MAC
ADJ 10 - Rewrite SRC A+DST J MAC



Software View

L3 Table: Design 2



IPv4/v6 Unicast FIB VPN / Prefix / Mask / Paths / Offset
1 / 10.1.2.0 / 24 / 4 / 1
1 / 10.1.3.0 / 24 / 1 / 5
3 / 10.1.2.0 / 24 / 2 / 6
3 / 10.1.3.0 / 24 / 2 / 6

Path Table
Path 1
Path 2
Path 3
Path 4
Path 1
Path 1
Path 2

Rewrite Information
ADJ 1 - Rewrite SRC A+DST A MAC
ADJ 2 - Rewrite SRC A+DST B MAC
ADJ 3 - Rewrite SRC A+DST C MAC
ADJ 4 - Rewrite SRC A+DST D MAC
ADJ 5 - Rewrite SRC A+DST E MAC
ADJ 6 - Rewrite SRC A+DST F MAC
ADJ 7 - Rewrite SRC A+DST G MAC
ADJ 8 - Rewrite SRC A+DST H MAC
ADJ 9 - Rewrite SRC A+DST I MAC
ADJ 10 - Rewrite SRC A+DST J MAC

Software View

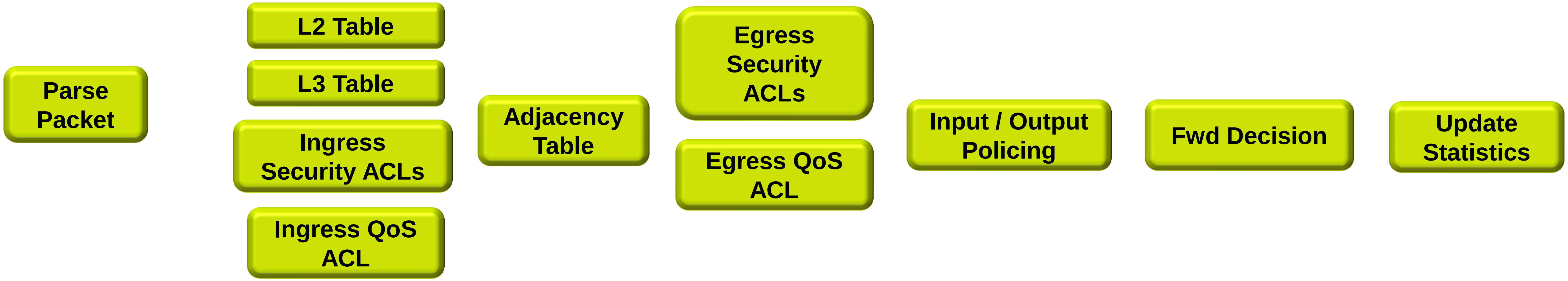
L2 Table / Host Table / FIB

Common Optimisation

- Hash tables take less space than TCAMs and Tries
- Instead of placing /32 or /128 entries for host entries into the FIB, place them into the hash table
- Common for the L2 table and the Host table to share the same memory
- Allows for the FIB Table to be smaller since it does not need to contain single path /32 and /128 entries

Forwarding Design

Design 1



Parse Packet

L2 Table

L3 Table

Ingress Security ACLs

Ingress QoS ACL

Adjacency Table

Egress Security ACLs

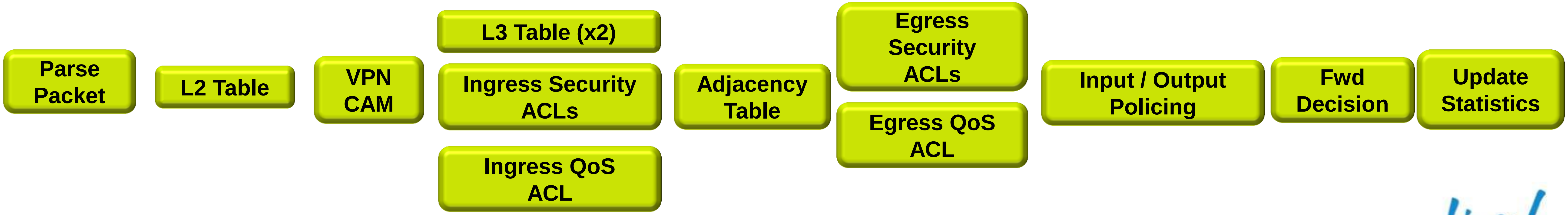
Egress QoS ACL

Input / Output Policing

Fwd Decision

Update Statistics

Design 2



Parse Packet

L2 Table

VPN CAM

L3 Table (x2)

Ingress Security ACLs

Ingress QoS ACL

Adjacency Table

Egress Security ACLs

Egress QoS ACL

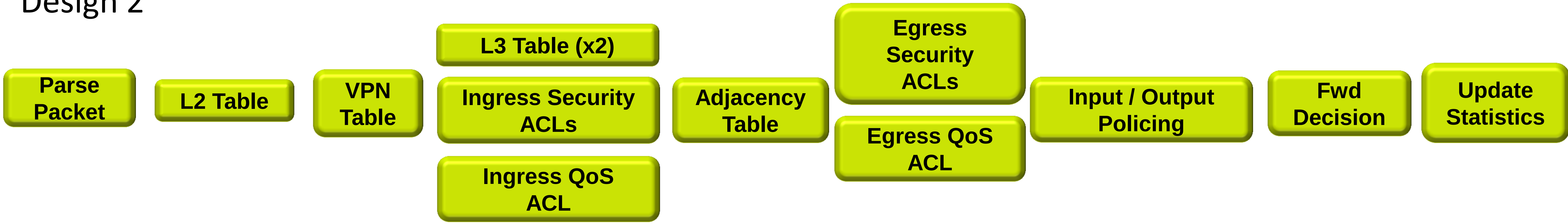
Input / Output Policing

Fwd Decision

Update Statistics

Forwarding Design

Design 2



Parse Packet

L2 Table

VPN Table

L3 Table (x2)

Ingress Security ACLs

Ingress QoS ACL

Adjacency Table

Egress Security ACLs

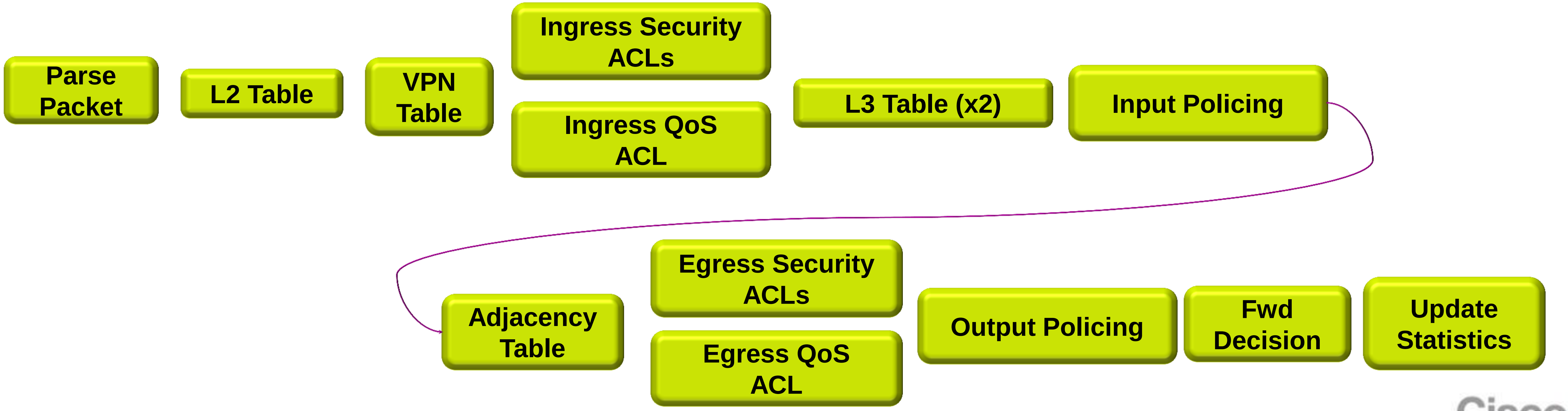
Egress QoS ACL

Input / Output Policing

Fwd Decision

Update Statistics

Design 3



Parse Packet

L2 Table

VPN Table

Ingress Security ACLs

Ingress QoS ACL

L3 Table (x2)

Input Policing

Adjacency Table

Egress Security ACLs

Egress QoS ACL

Output Policing

Fwd Decision

Update Statistics

References

- Network Algorithmics,: An Interdisciplinary Approach to Designing Fast Networked Devices George Varghese
- Art of Computer Programming Vol 1-4, Donald E. Knuth
- Introduction to Algorithms, Third Edition
  Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein
- IEEE SIGCOMM Papers

ASIC Engineering

ASICs vs FPGAs

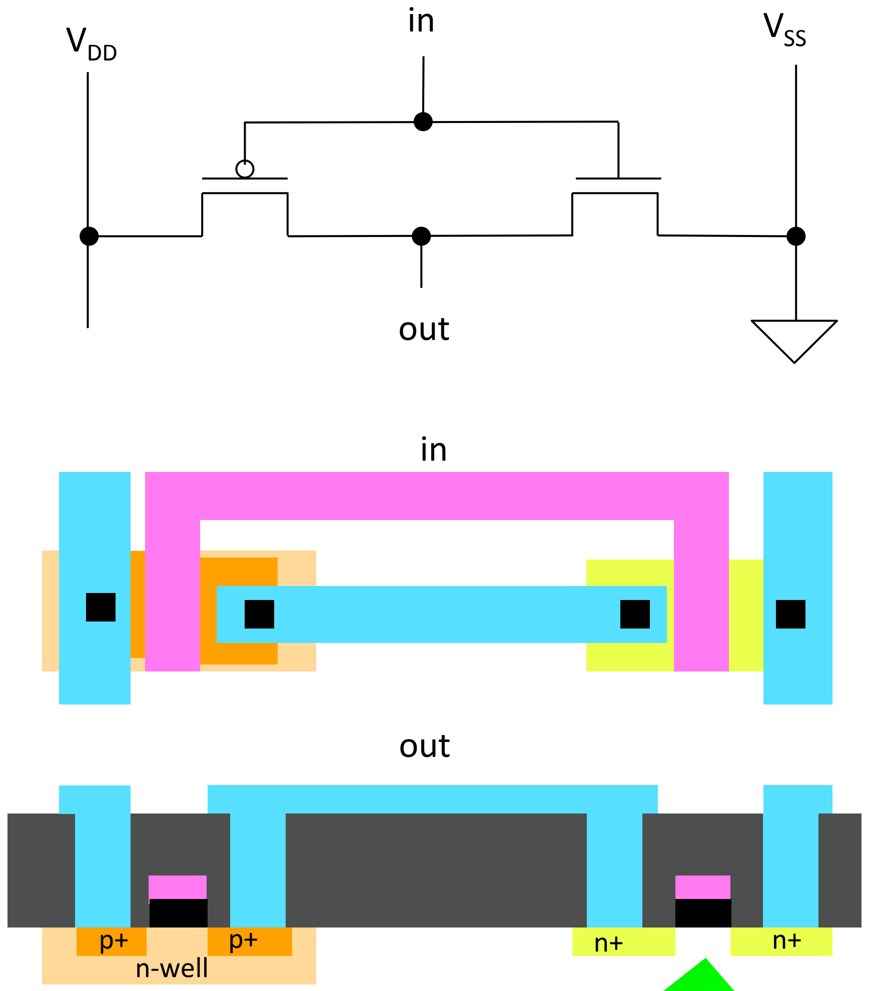
ASIC - Application Specific Integrated Circuit

- A finished IC which is built to the exact specification & functionality of the customer
- Can make optimal use of the underlying silicon circuits
- Low part cost, High upfront investment
- Significant development time

FPGA (EPLD) Field Programmable Gate Array

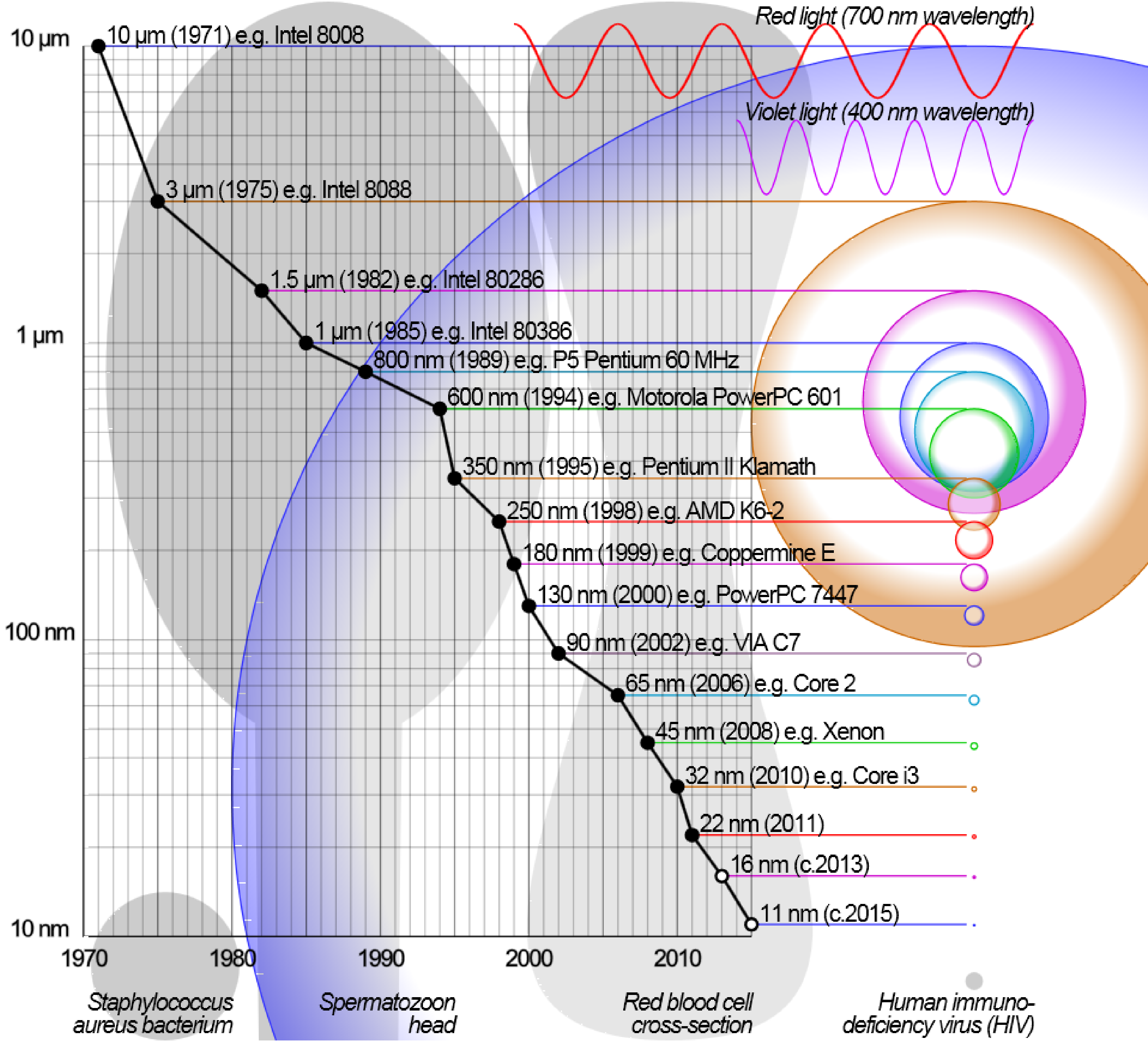
- An IC that can be configured with the required functionality after it is installed into a target system
- Flexibility vs. sub-optimal use of underlying silicon circuits
- Higher part cost
- Shorter development time
- Main players: Xilinx, Altera

CMOS



"Feature size"

This dimension is what Moore's Law is all about!

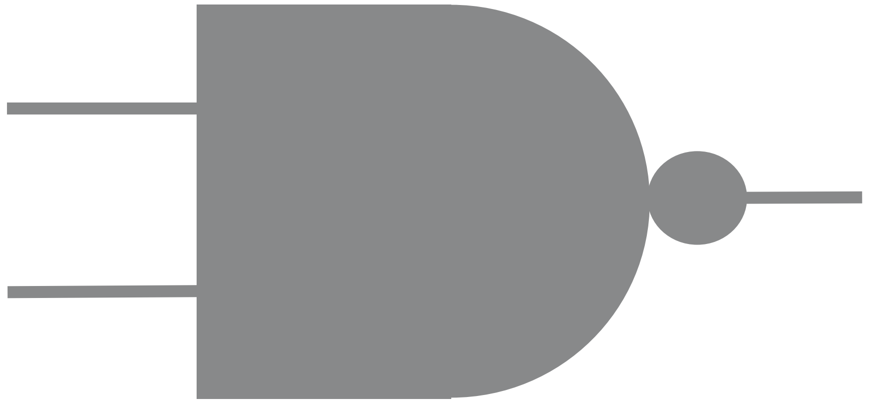


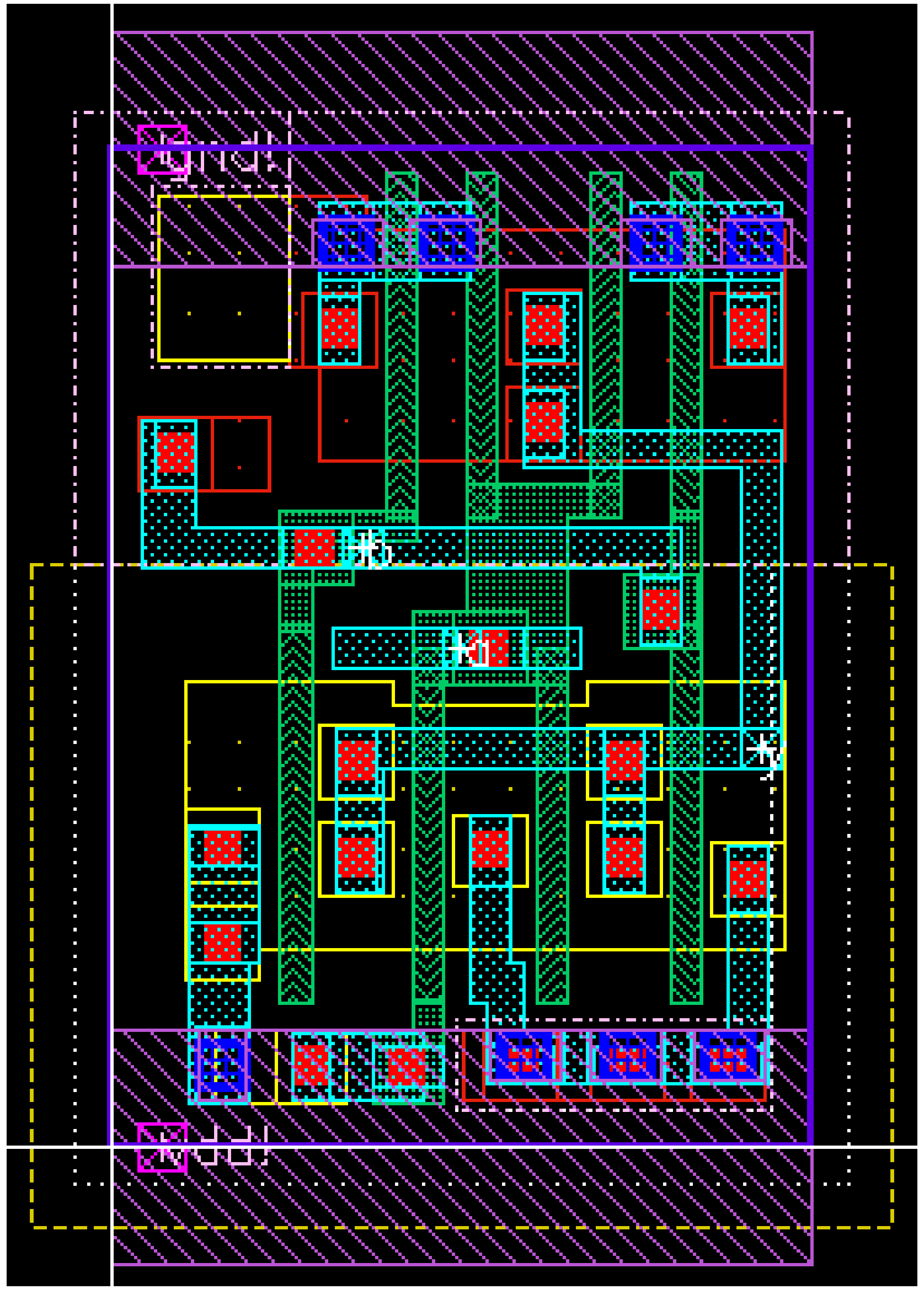
http://en.wikipedia.org/wiki/Semiconductor_device_fabrication

1.6 X Increase in usable gates between process nodes

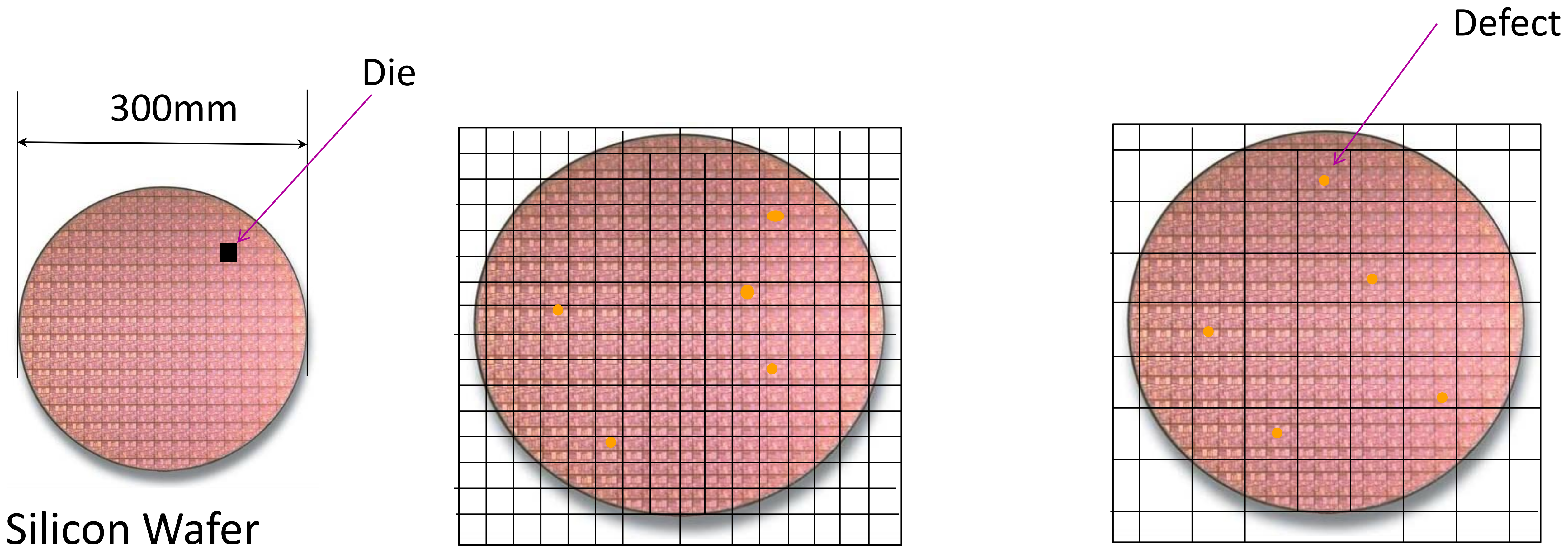
Gates

```
module nand2(a,b,c)
  input a,b;
  ouput c;
begin
       c <= !(a & b);
end
```



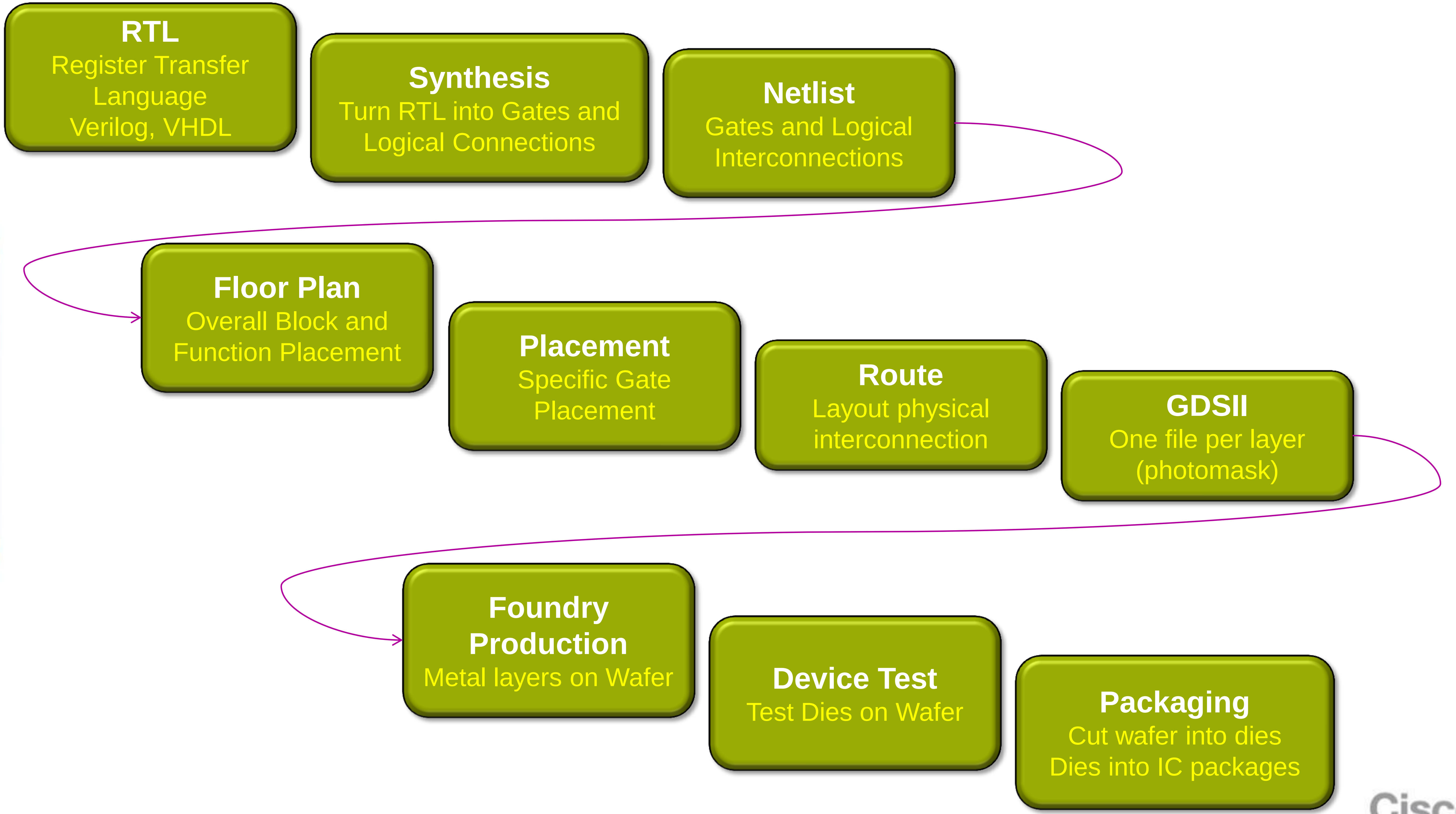


Why is Die Size Important?



With same number of defects per wafer,
smaller die size results in higher yield per wafer

Integrated Circuit Production



RTL
Register Transfer Language
Verilog, VHDL

Synthesis
Turn RTL into Gates and Logical Connections

Netlist
Gates and Logical Interconnections

Floor Plan
Overall Block and Function Placement

Placement
Specific Gate Placement

Route
Layout physical interconnection

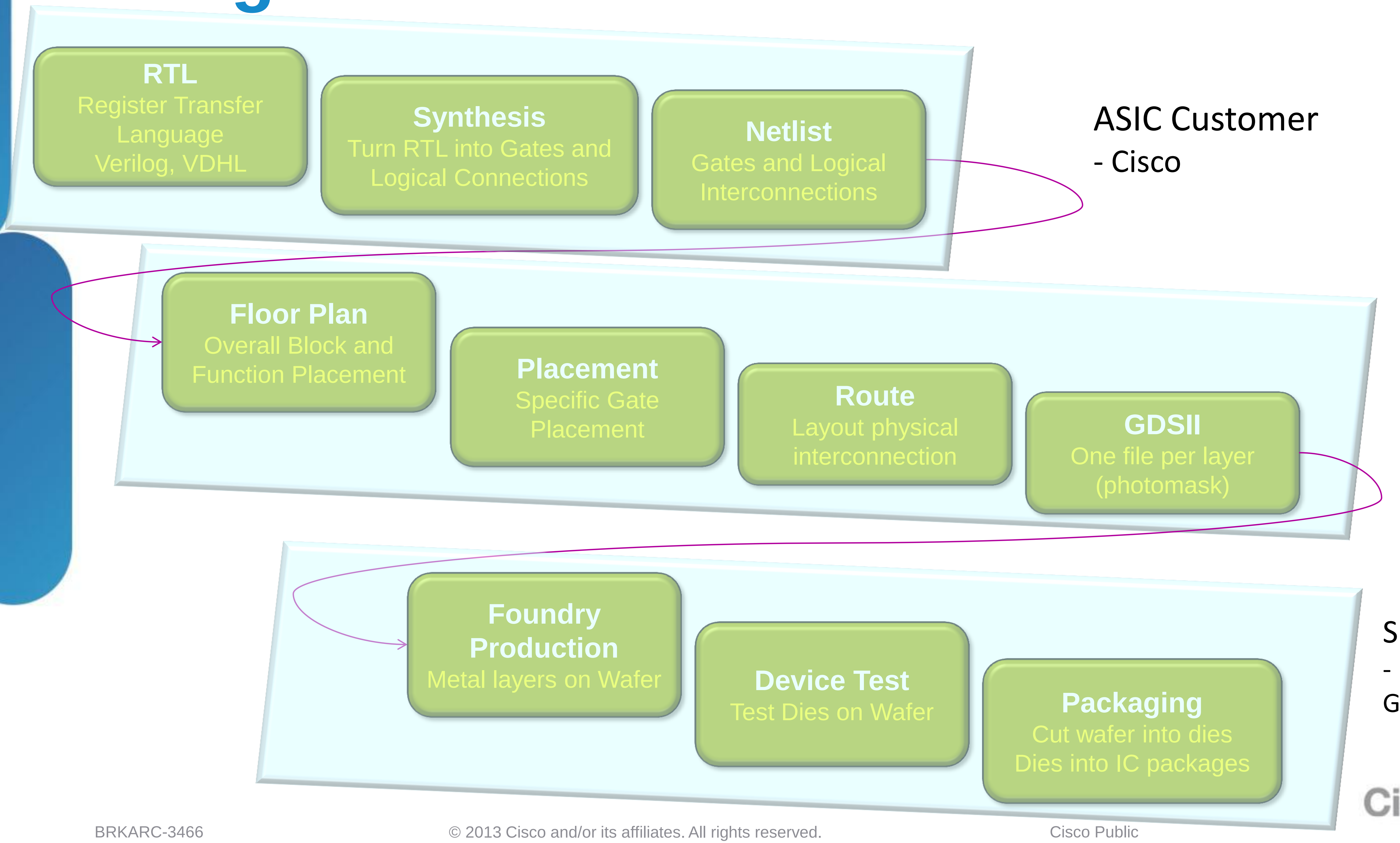
GDSII
One file per layer (photomask)

Foundry Production
Metal layers on Wafer

Device Test
Test Dies on Wafer

Packaging
Cut wafer into dies
Dies into IC packages

Integrated Circuit Production



RTL
Register Transfer Language
Verilog, VDHL

Synthesis
Turn RTL into Gates and Logical Connections

Netlist
Gates and Logical Interconnections

Floor Plan
Overall Block and Function Placement

Placement
Specific Gate Placement

Route
Layout physical interconnection

GDSII
One file per layer (photomask)

Foundry Production
Metal layers on Wafer

Device Test
Test Dies on Wafer

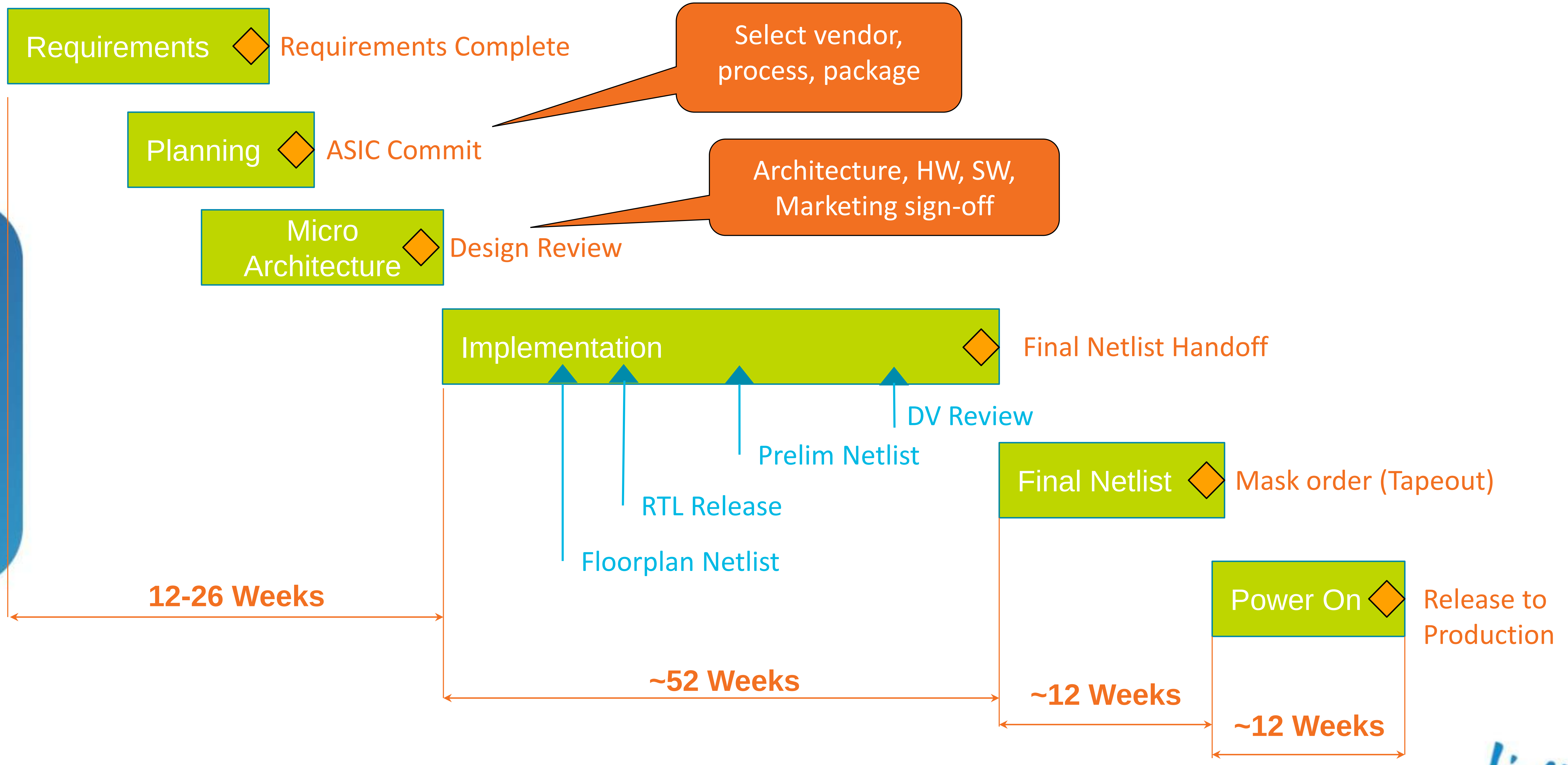
Packaging
Cut wafer into dies
Dies into IC packages

ASIC Customer
- Cisco

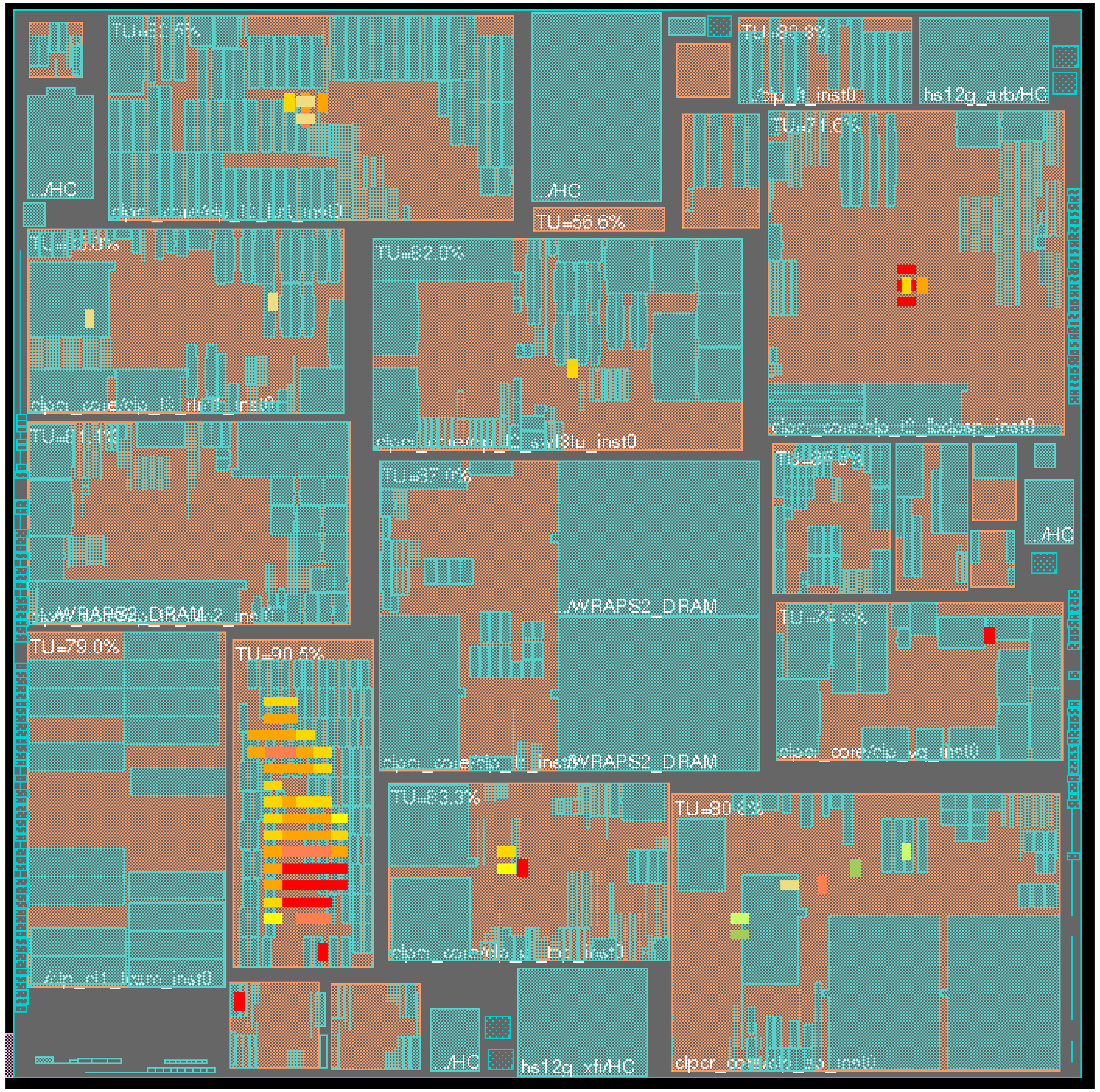
ASIC Vendor
- Avago, IBM, TI, ST Micro
- COT - Cisco

Silicon Foundry
- IBM, TSMC, Global Foundries

ASIC Design Process



F2/F2E ASIC - Clipper



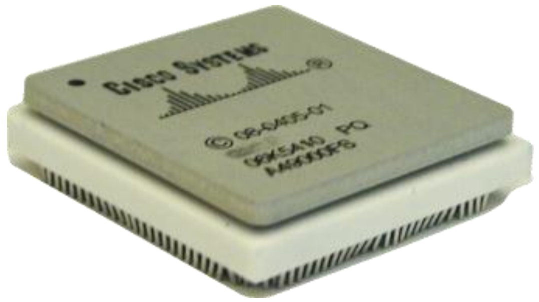
	F2	F2E
Technology	IBM Cu-65	IBM Cu-45
Die Size	18.0x18.3mm	12.28x12.28mm
Total SRAM	33.3Mb	33.3Mb
Total eDRAM	134Mb	134Mb
Total TCAM	2.94Mb	2.94Mb
Register Array	1.34Mb	1.34Mb
Logic Gates	45M	45M
Signal Pin	186	186
Package IO	840	840

Memory and Packet Corruption Protection

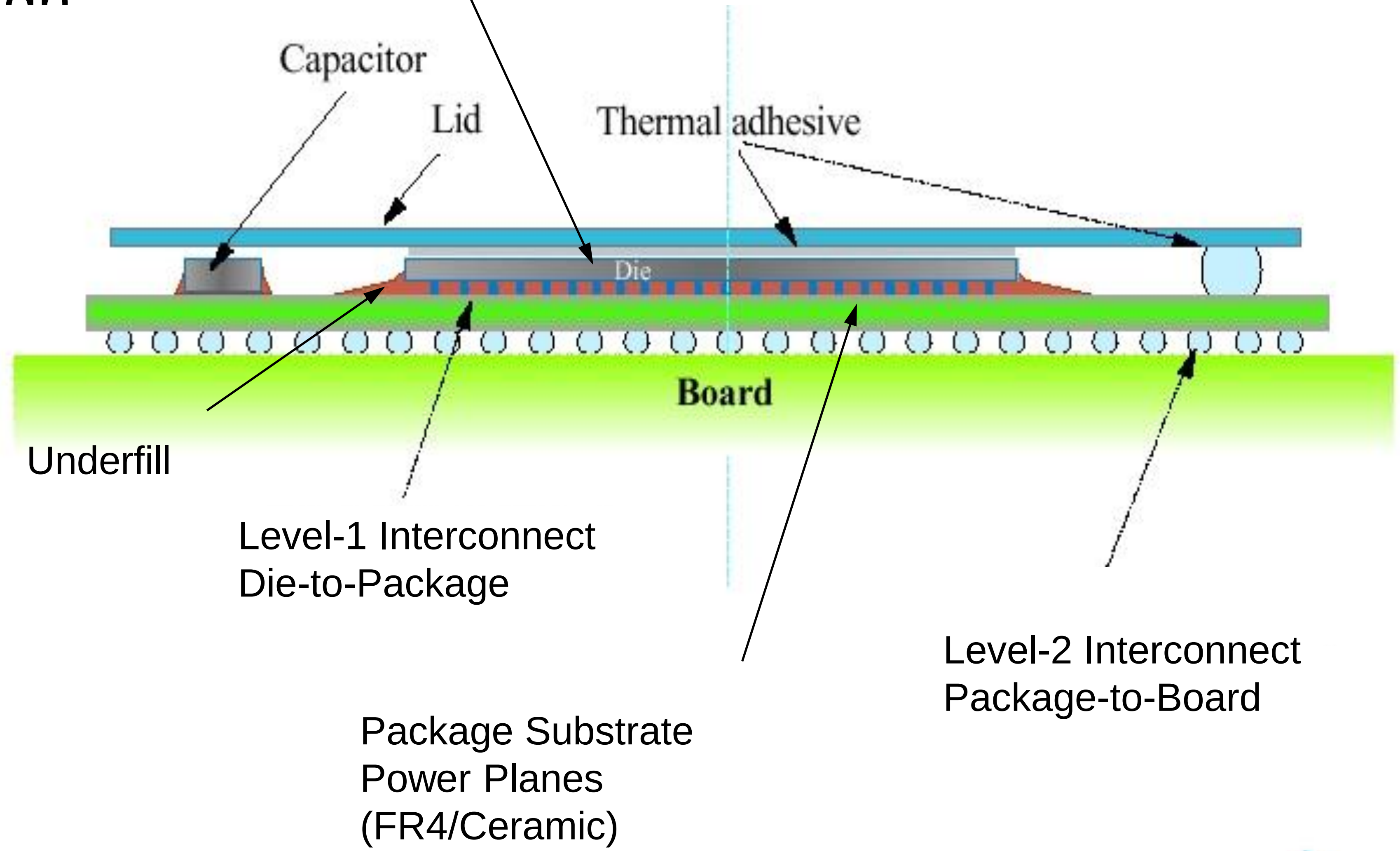
- No ECC or Parity – no way to determine if a software or hardware problem
- Parity – will detect single bit errors
- ECC – will detect 2 bit errors, and correct single bit
- Parity and ECC apply to a word (32 or 64 bits)
- CRC – Detect if a set of bytes (normally a packet) has been corrupted

ASIC Packaging

- Electrical parasitics of the chip package are critical
- Impacts electrical properties of high speed signals
- Manufacturing tolerances constra minimum ball pitch
- Limit to number of available signal I/O pins



Silicon Die



Underfill

Level-1 Interconnect
Die-to-Package

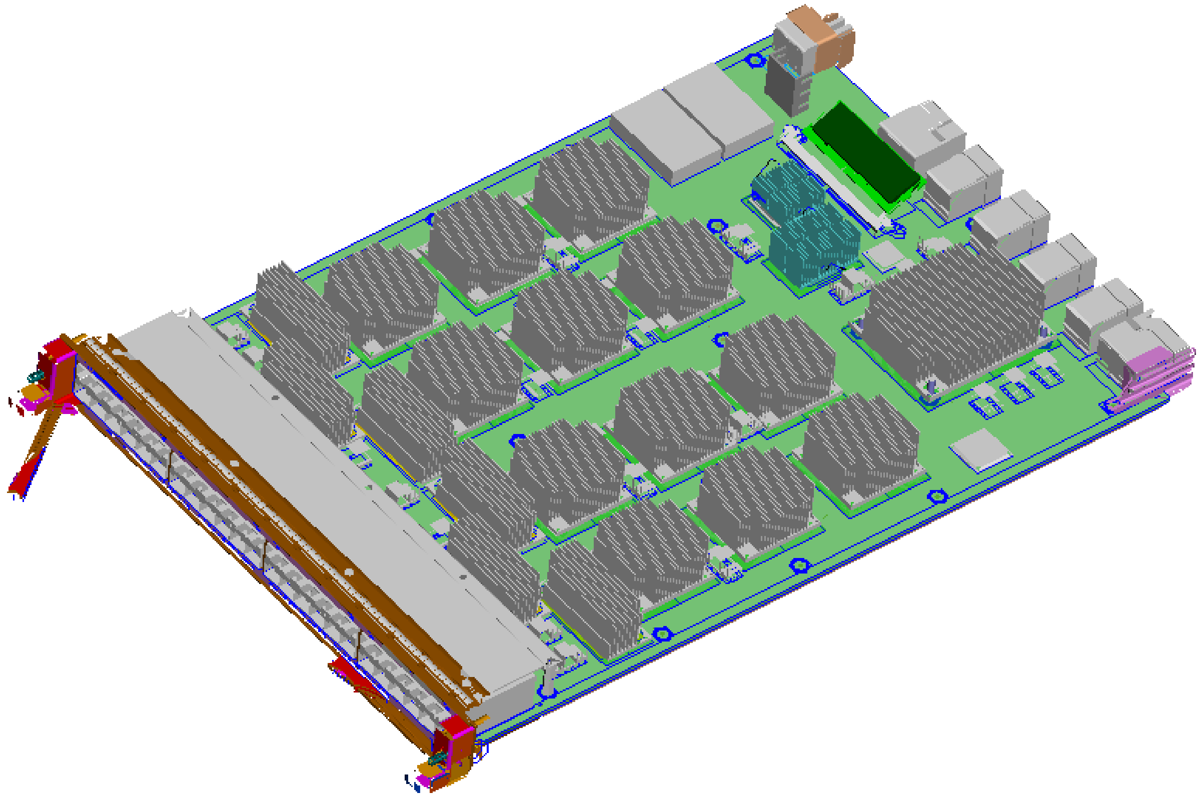
Package Substrate
Power Planes
(FR4/Ceramic)

Level-2 Interconnect
Package-to-Board

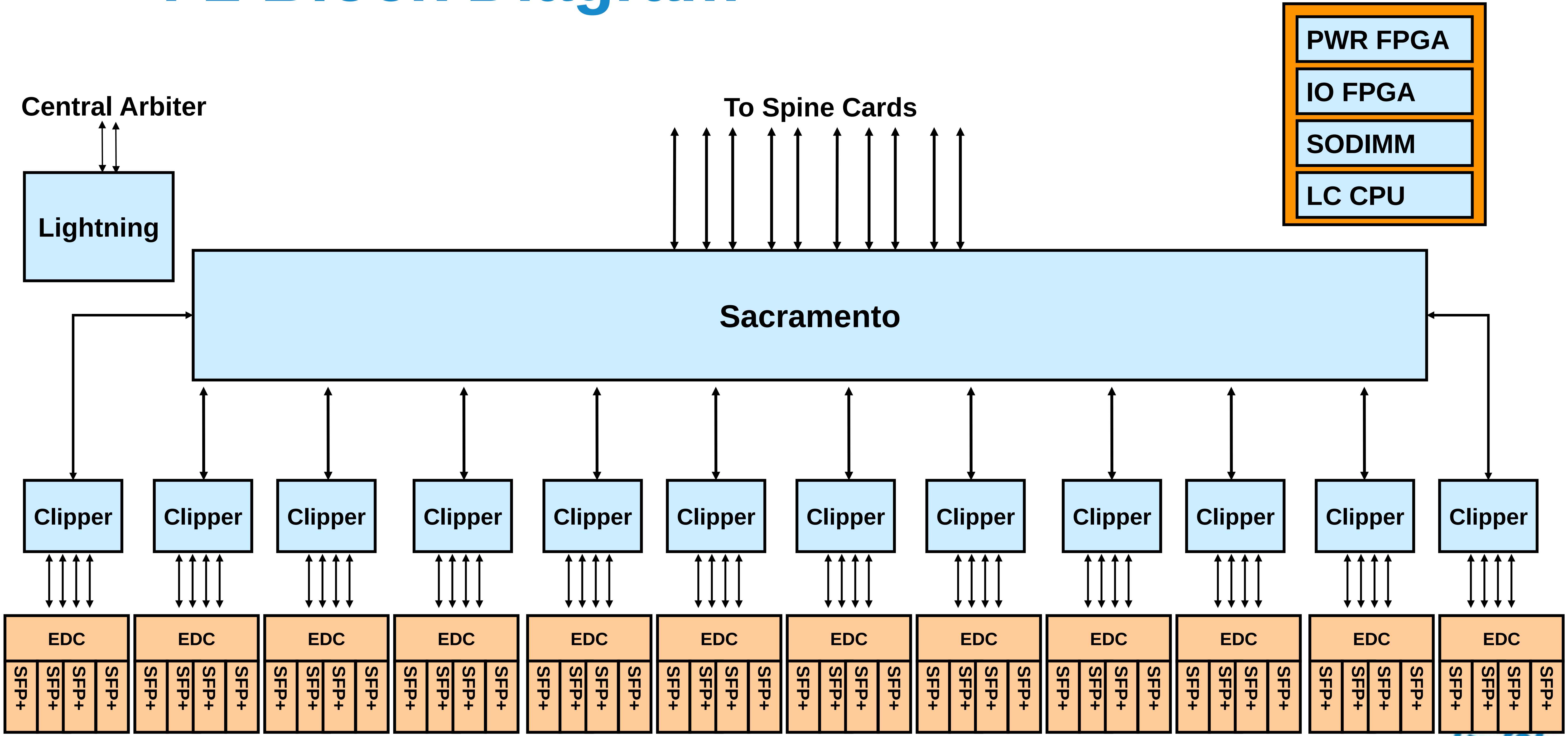
References

- Indistinguishable From Magic: Manufacturing Modern Computer Chips http://www.youtube.com/watch?v=NGFhc8R_uO4

Hardware Engineering

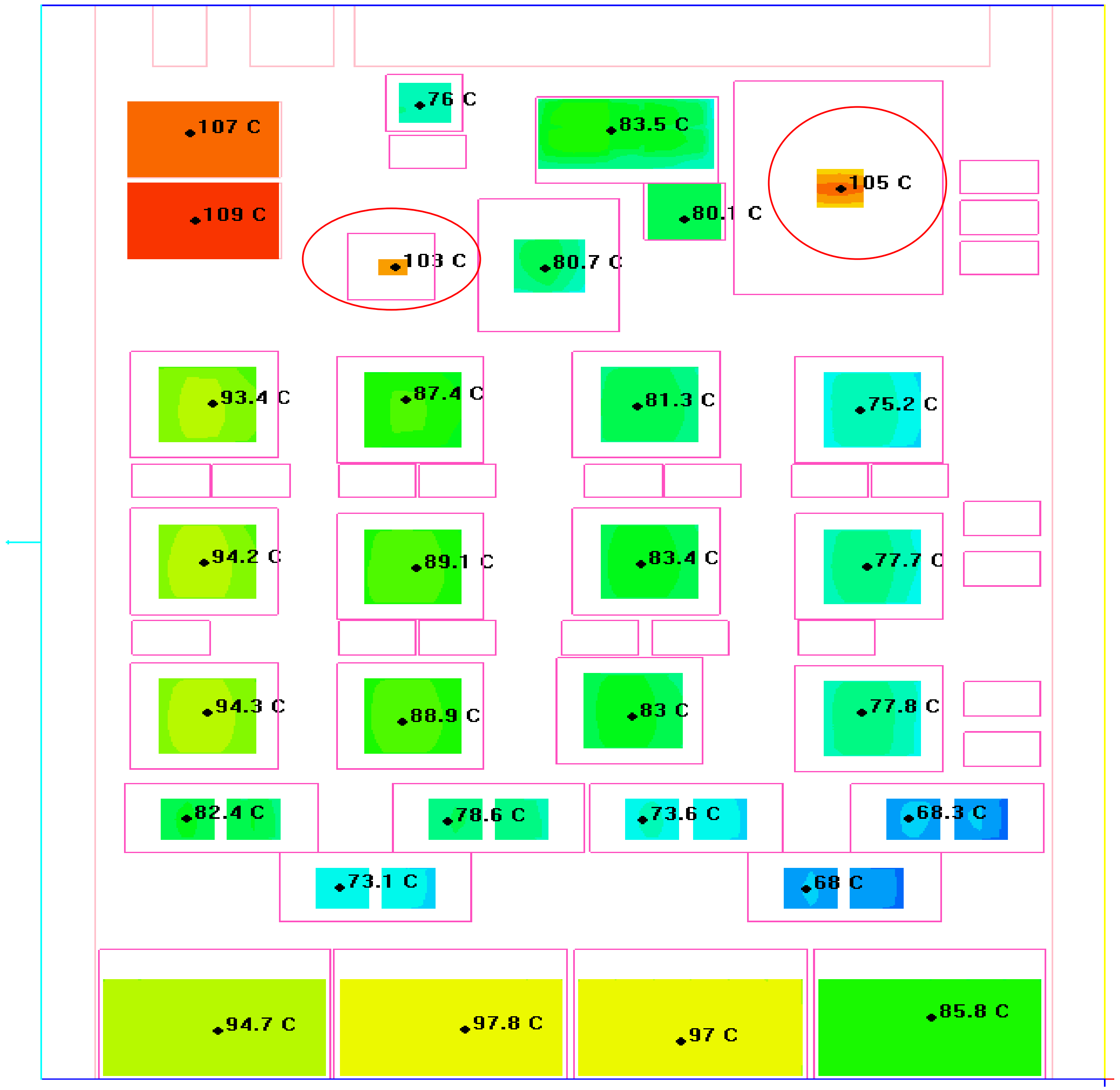


F2 Block Diagram

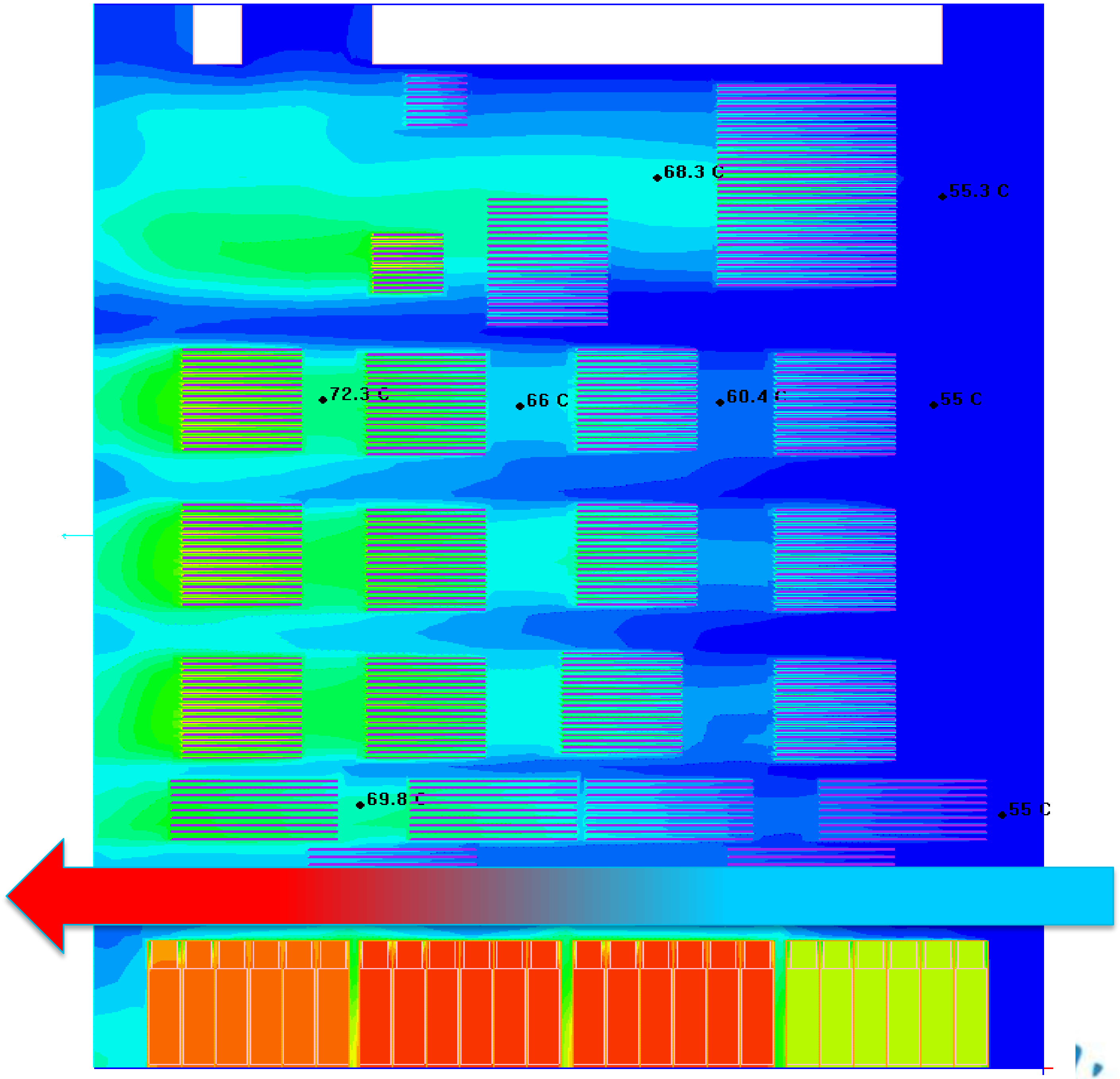


Thermal Modeling

Component Case Temperatures



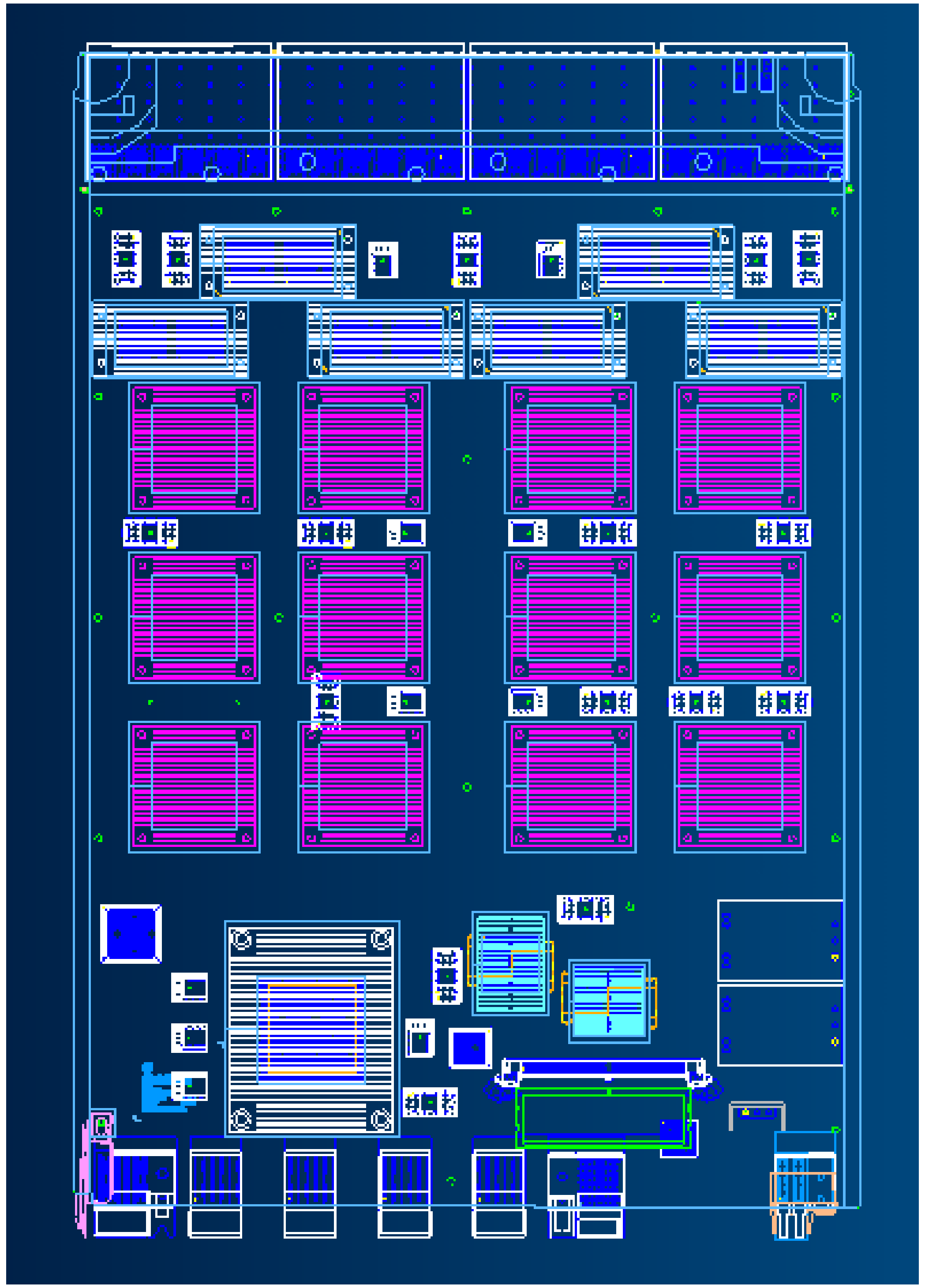
Temperature Contours

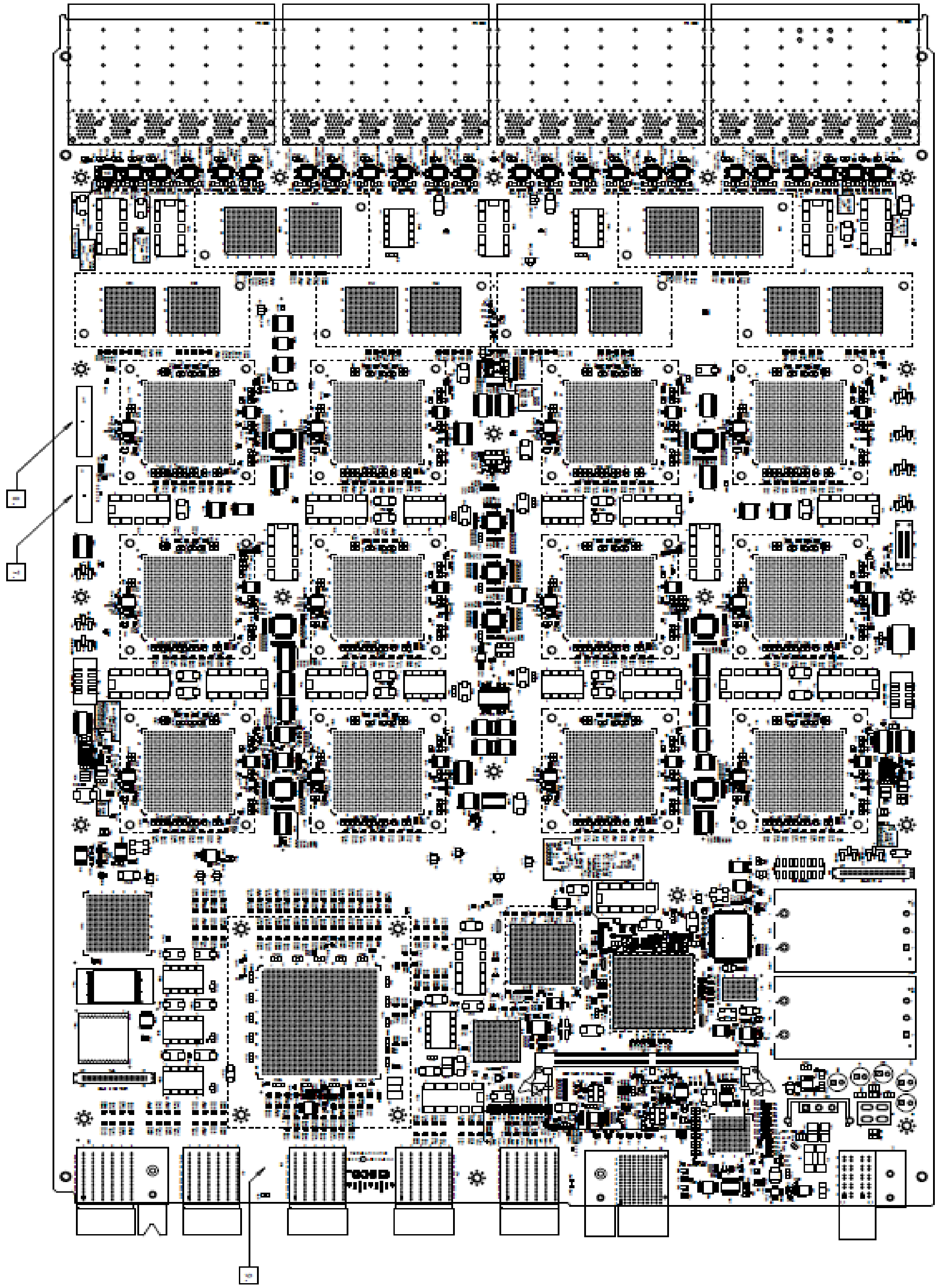


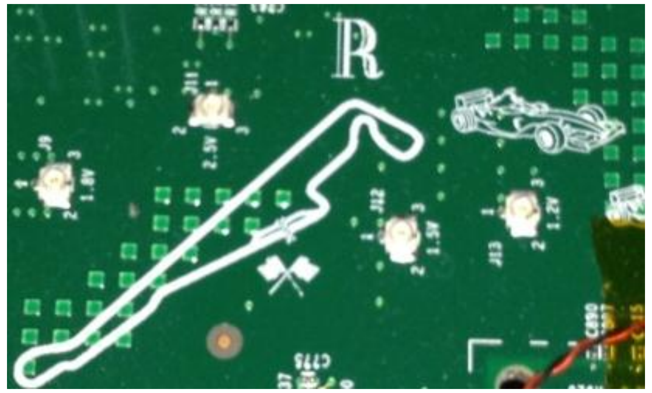
55C Inlet Temperature

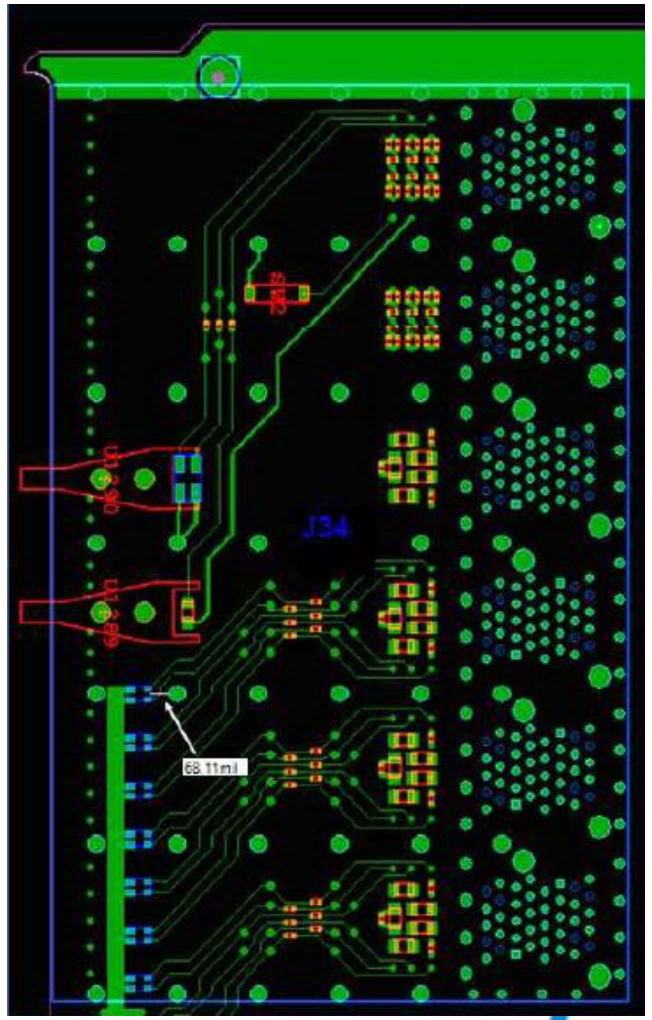
Electrical / Mechanical Layout

20 Layers









EDVT

Electronic Design Validation Test

- All tests performed using offline diagnostics and again with NXOS
- On-board power supplies have voltages margined to +5% & -5%
- Temperature testing occurs while
- Soaking for 12 hours at 55^{o} C and -5^{o} C
- Ramping between extremes at 1^{o} C per minute
- Power cycle testing occurs during 12-hour soak

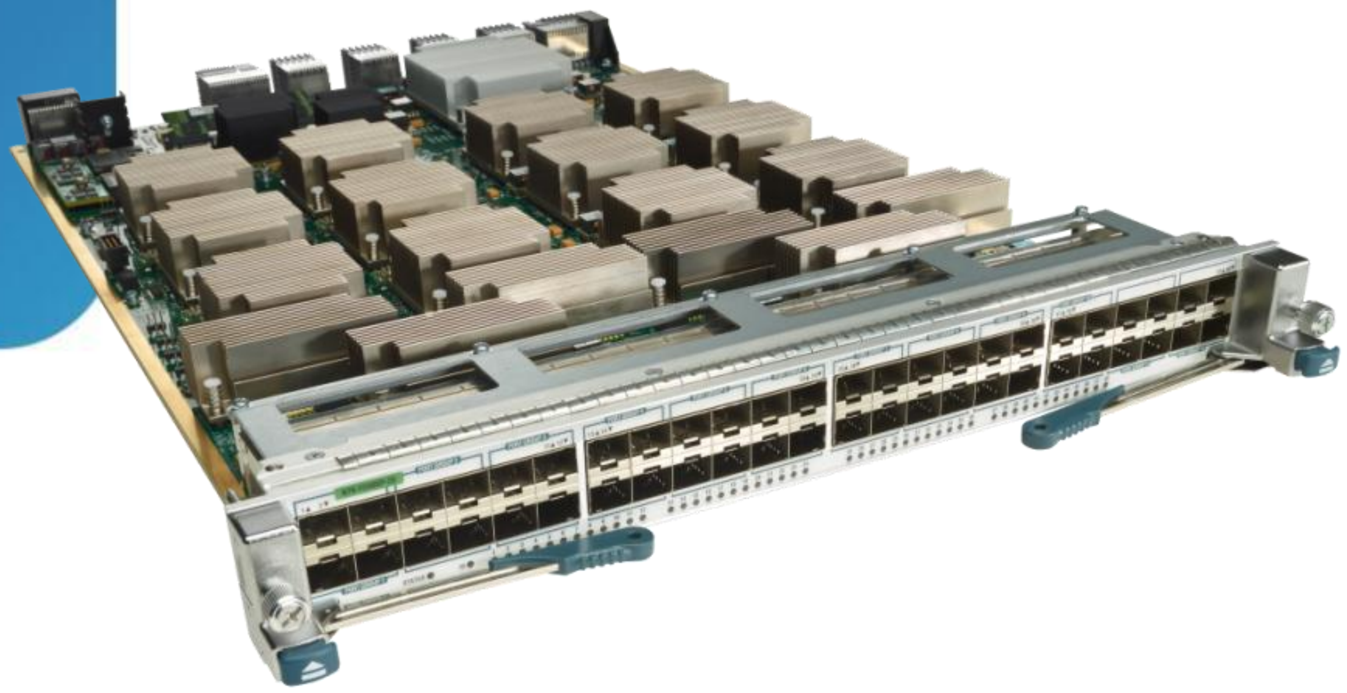
RDT

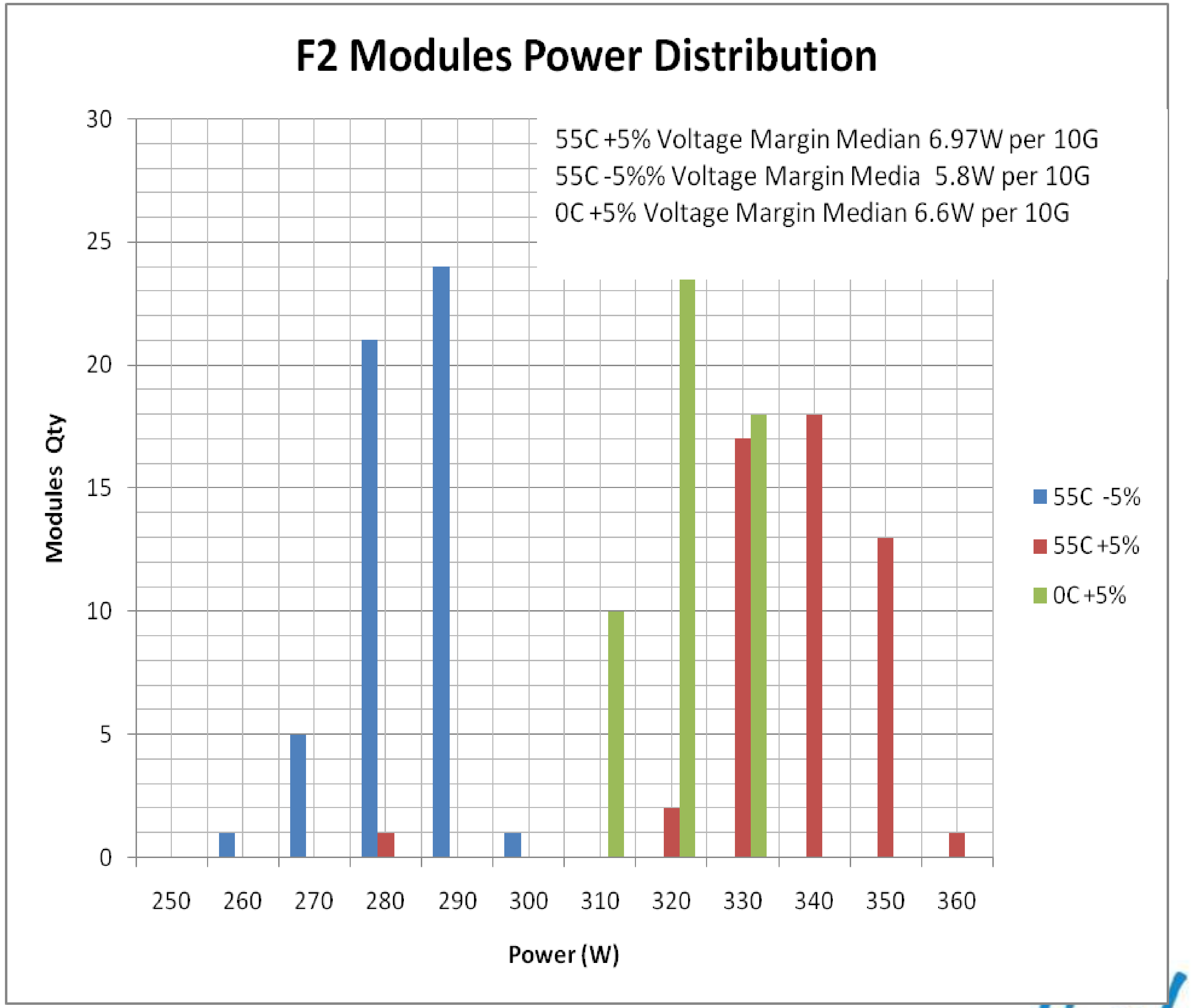
Reliability Demonstration Test

- The Reliability Demonstration Test (RDT) is Cisco’s approach to verifying the stated reliability of a product prior to production release.
- The reliability to be demonstrated is the product’s MTBF (Mean Time Between Failure).
- RDT replicates the end user operating environment and application through accelerated test time. It is expected that all hardware features are exercised in RDT.
- All new products including systems and boards are subject to RDT.

Power Consumption

- Skew Parts
- Data Sheet
  - – Typical 340W
  - – Maximum 450W





F2 Modules Power Distribution

55C +5% Voltage Margin Median 6.97W per 10G
55C -5%% Voltage Margin Media 5.8W per 10G
0C +5% Voltage Margin Median 6.6W per 10G

Generic Online Diagnostics

Generic Online Diagnostics provide a diagnostic framework for detecting hardware faults and verifying the health of hardware components throughout the chassis.

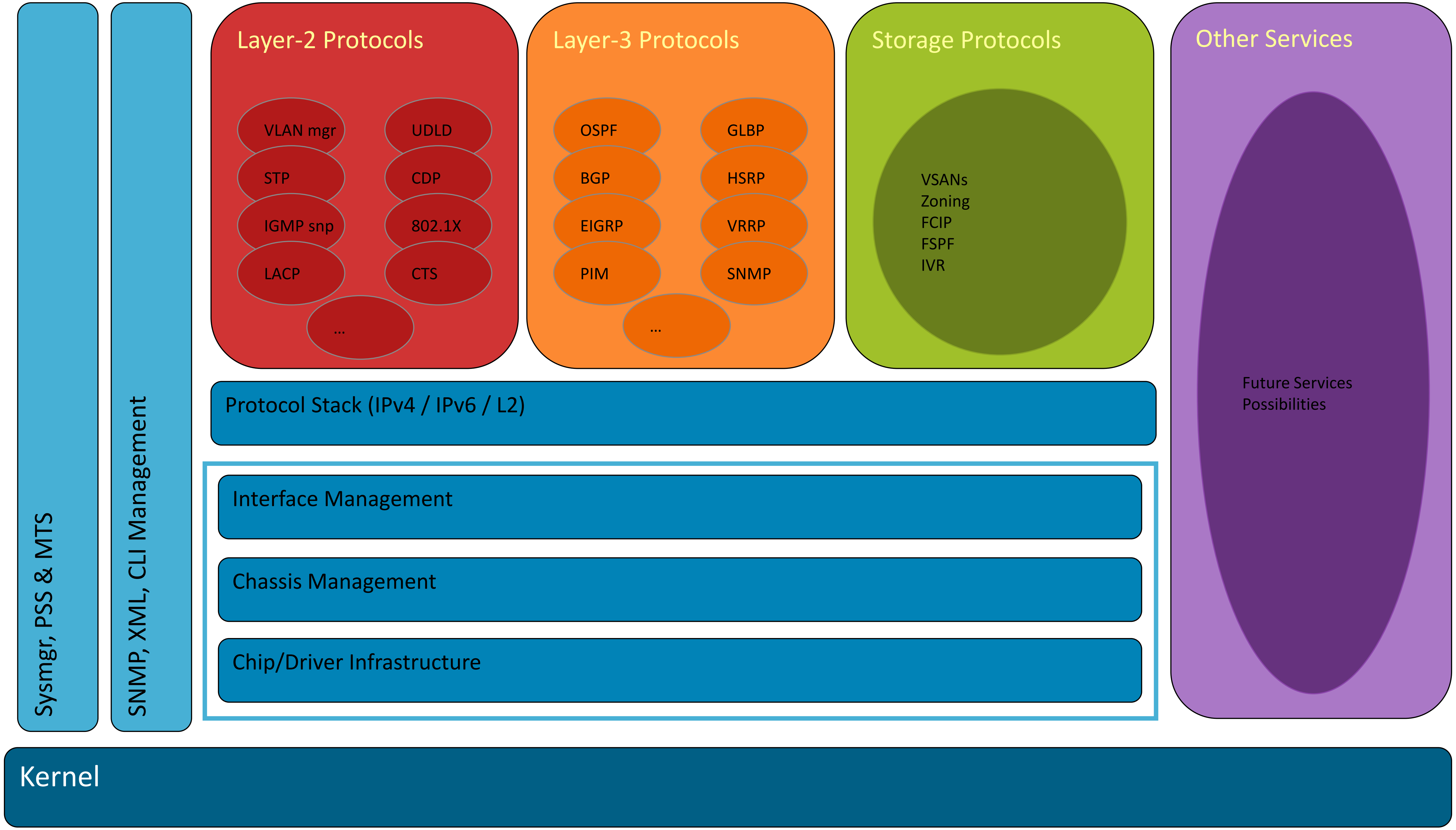
Diagnostics run during system Boot-Up, after OIR, On-Demand using the CLI, or as Health Checks in the background.

Problem Areas:

- Hardware Components (ASICs)
- Interfaces (Ethernet, SFP+, etc…)
- Connecters (loose connectors, bent pins, etc…)
- Memory Failure (Failure over time)
- Solder Joints

Software Engineering

NXOS Architecture



Multi-threaded

Scalability with SMP and multi-core CPUs

Faster Route Re-convergence

Lower mean-time-to-recovery

Real-Time

Real-Time preemptive scheduling

System operational when CPU is 100%

Modularity

Most of the features are conditional

Can be enabled/disabled independently

Maximises efficiency

Minimises resources utilisation

Separation Control Plane and Data Plane

No “software forwarding feature”

Fully distributed hardware forwarding

Line Card Offloading

Offload to line card CPUs

Scales with # of line cards

Optimal hardware programming

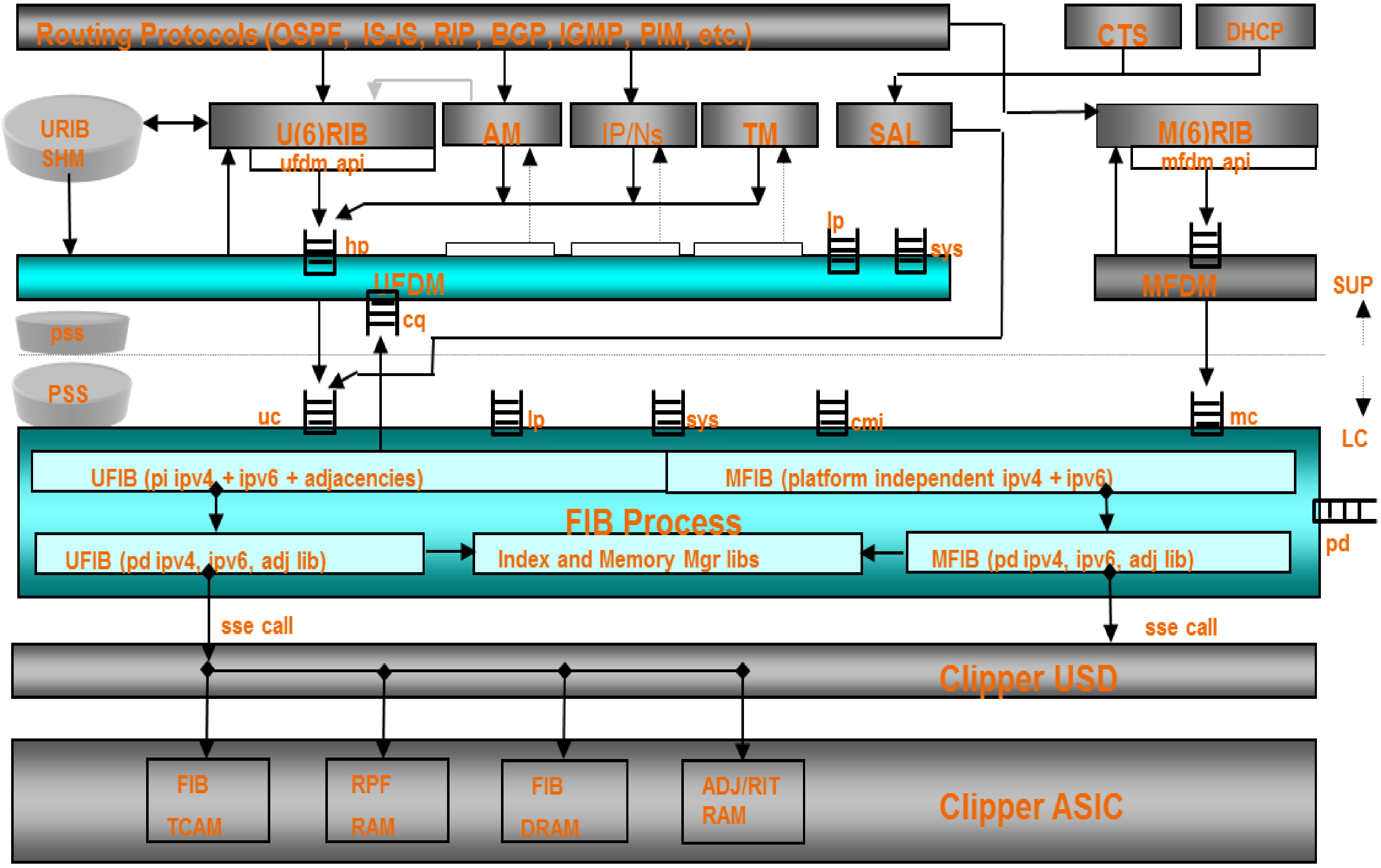
Software Engineering

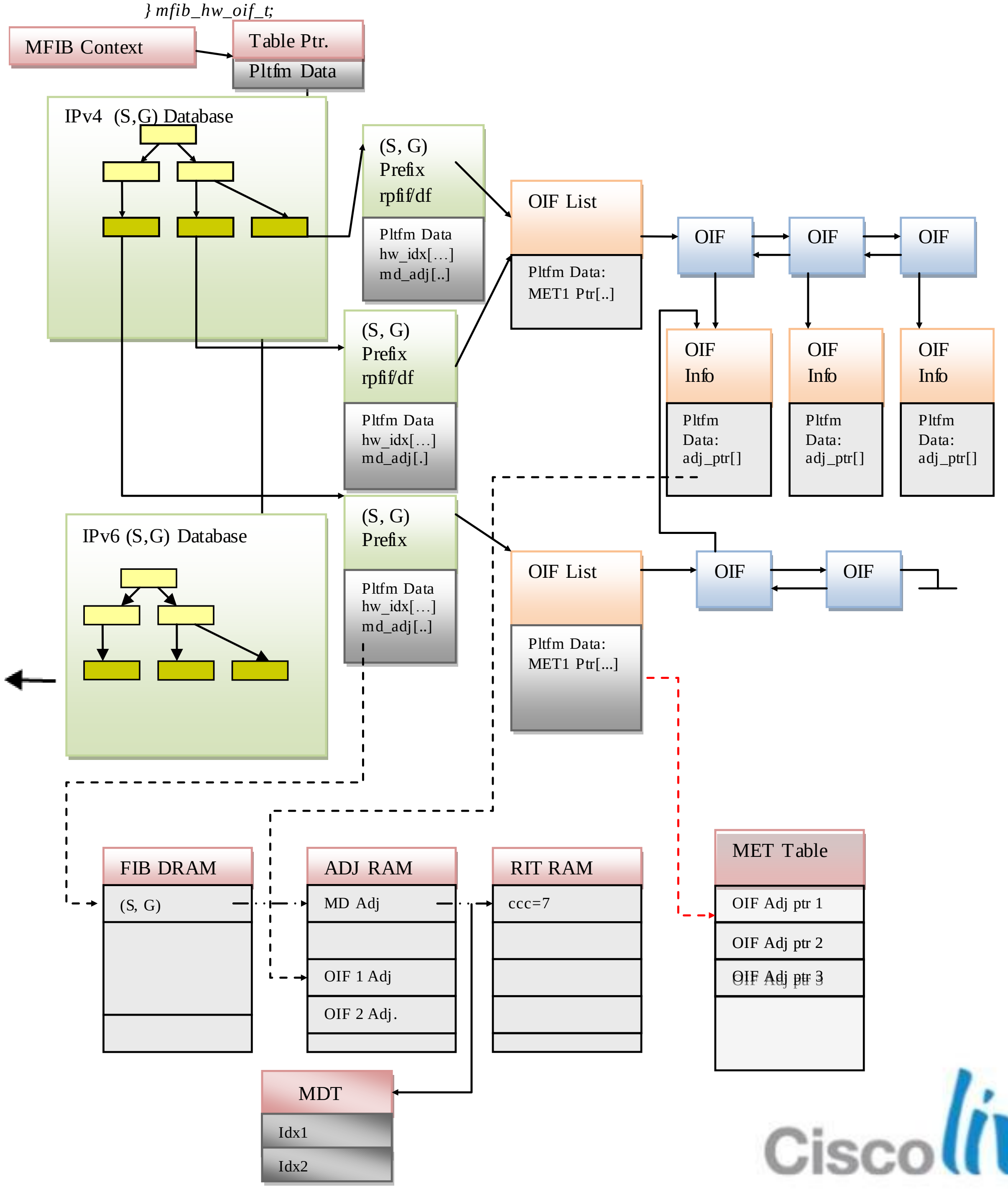
SW Functional Spec

SW Design Spec

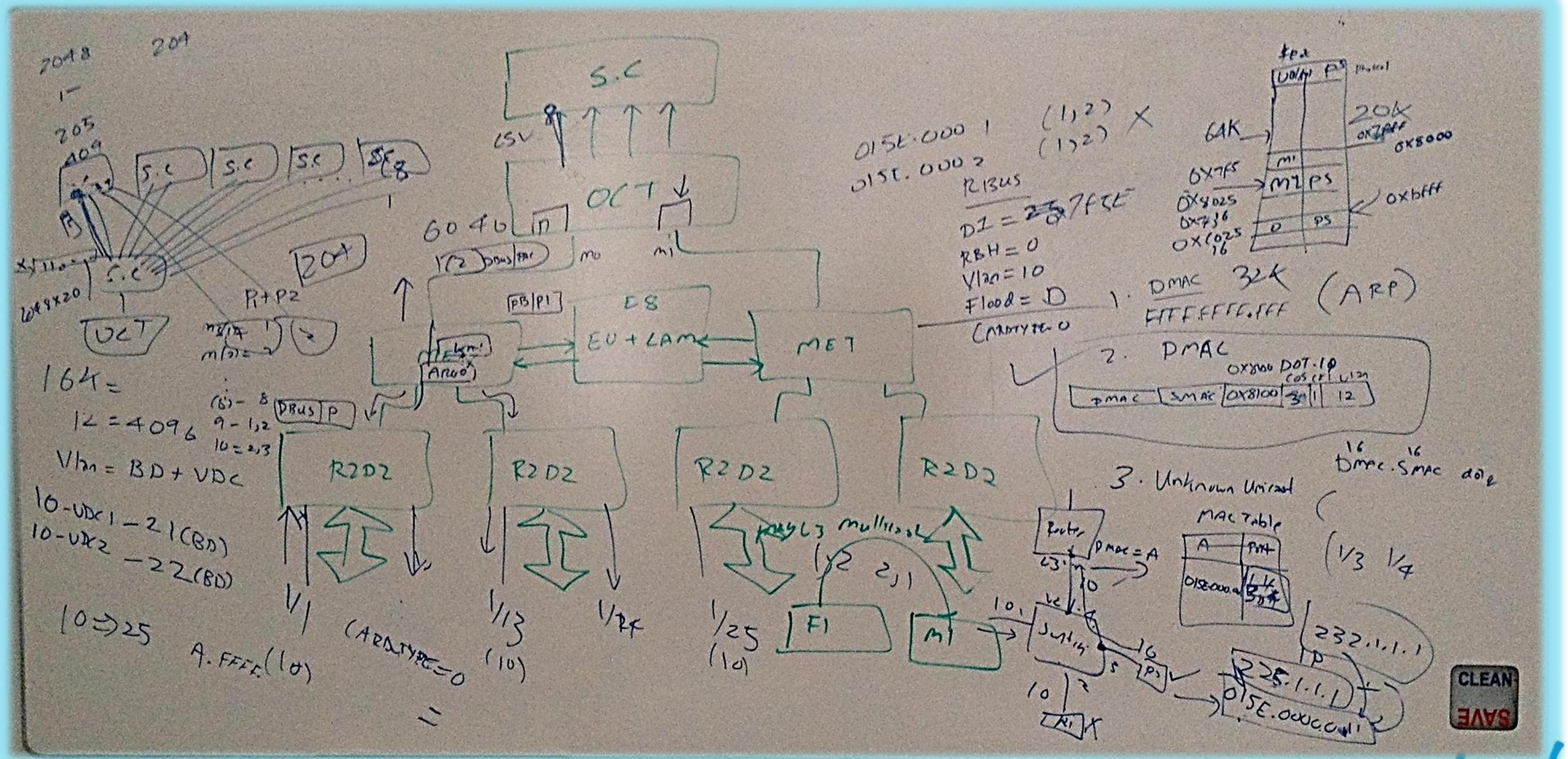
Unit Test Plan

Unit Integration Plan


Routing Protocols (OSPF, IS-IS, RIP, BGP, IGMP, PIM, etc.)
CTS
DHCP
URIB
SHM
U(6)RIB
ufdm_api
AM
IP/Ns
TM
SAL
M(6)RIB
mfdm_api
hp
lp
sys
UFDM
MFDM
SUP
LC
pss
PSS
cq
uc
lp
sys
cmi
mc
UFIB (pi ipv4 + ipv6 + adjacencies)
MFIB (platform independent ipv4 + ipv6)
FIB Process
UFIB (pd ipv4, ipv6, adj lib)
Index and Memory Mgr libs
MFIB (pd ipv4, ipv6, adj lib)
pd
sse call
sse call
Clipper USD
FIB
TCAM
RPF
RAM
FIB
DRAM
ADJ/RIT
RAM
Clipper ASIC



Design Review



Development Test

Master Test Plans | Functional Test Plans | Automation | Regression | FCS

- Testing of completed integrated feature
- Test for interactions with other features and functions
- Test for interoperability with Cisco and 3rd party devices
- Build scripts to automate testing so is repeatable on future releases

First Customer Ship

Product Requirements Document

ASIC

Requirements ◆ Pl ◆ lementation ◆ Final Netlist ◆ Power On

Hardware

HW Des

Me

Elect

Manufactu

P1 ◆ P2 ◆ A-0

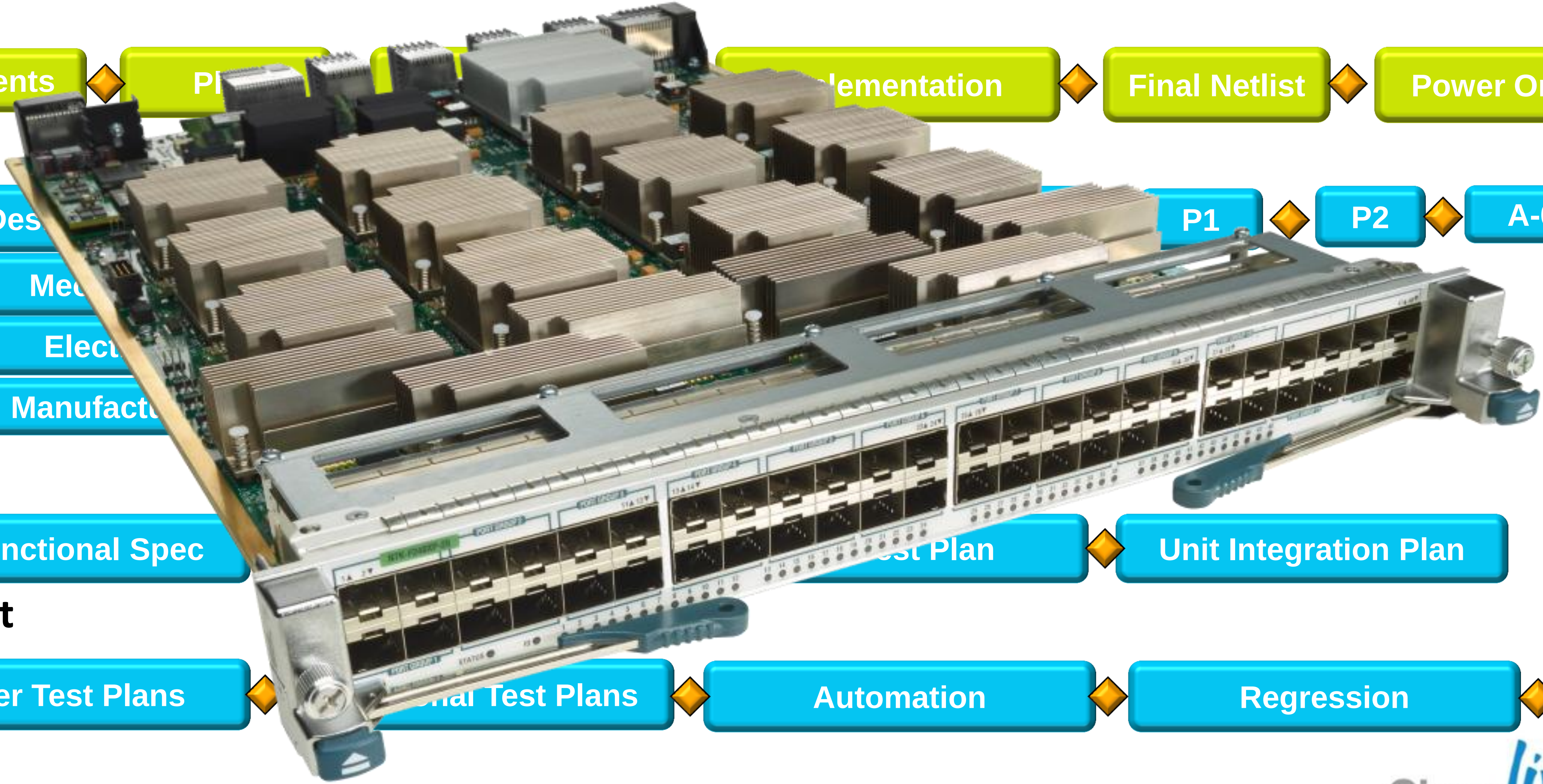
Software

SW Functional Spec

est Plan ◆ Unit Integration Plan

Software Test

Master Test Plans ◆ nal Test Plans ◆ Automation ◆ Regression ◆ FCS



Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2013 Polo Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 8 March 12:00pm-2:00pm

Cisco live! 365

Don't forget to activate your Cisco Live 365 account for access to all session material, communities, and on-demand and live activities throughout the year. Log into your Cisco Live portal and click the "Enter Cisco Live 365" button. www.ciscoliveaustralia.com/portal/login.ww