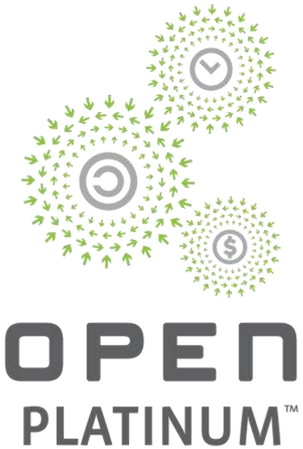
Open. Together.

OCP
REGIONAL
SUMMIT

SONiC – Reliability, Manageability and Extensibility

Guohan Lu| Xin Liu
Azure Networking, Microsoft

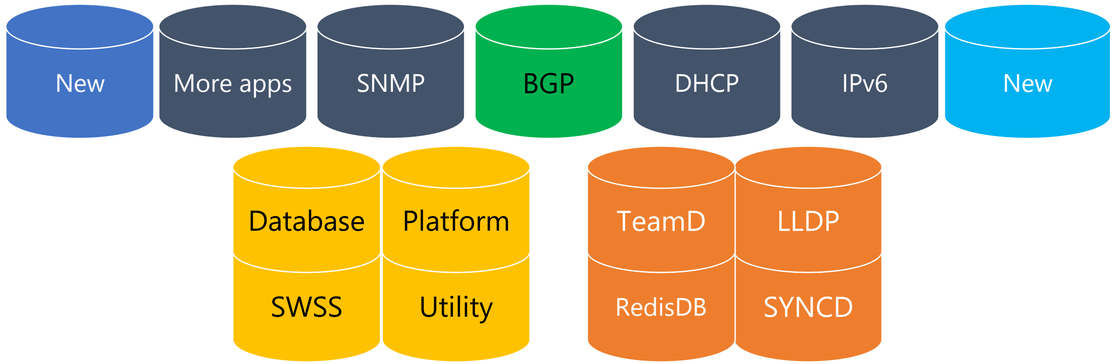
Ben Gale
Broadcom Inc



SONiC Software for Open Networking in the Cloud

configuration and management tools

Jenkins ANSIBLE kubernetes docker SWARM puppet CHEF 1ST party



SONiC

Linux

Switch Abstraction Interface (SAI)

Powering data center ToR/Leaf → Powering bare metal service → Powering AI/gaming service

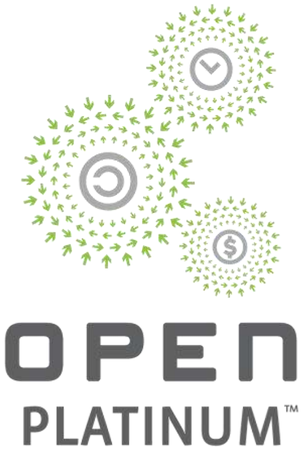
More contributors and hardware support → More industry adoption → Commercial support

➢ Linux ➢ Basic L2/L3 ➢ Containerized ➢ Redis DB	➢ RDMA/QoS ➢ IPv6 ➢ Mgmt. via Swarm ➢ Fast Reboot(<30s)	➢ Streaming Telemetry ➢ Config DB ➢ Support Virtualization ➢ Warm Reboot (<1s)	➢ Richer Features ➢ Advanced Mgmt ➢ Stringent Tests ➢ Development Tools
➢ 40G ➢ 5 platforms	➢ 100G ➢ 16 platforms	➢ ARM based & Lower end ➢ 31 platforms	➢ Chassis Support ➢ 92 platforms
➢ ASIC BRCM: Trident 2 MLNX: Spectrum Cavium: Xpliant Centec: Goldengate	➢ ASIC BRCM: Tomahawk/ Tomahawk2 Marvell: Prestera Barefoot: Tofino	➢ ASIC Nephos: Taurus BRCM: TD2/TH3, Helix4 Cisco: Lacrosse	➢ ASIC BRCM: DNX Innovium: Teralynx Marvell: Falcon MLNX: Spectrum II


ARISTA
BAREFOOT NETWORKS
BROADCOM
Alibaba Group
apstra
docker
centec networks
Celestica
CANONICAL
CISCO
Mellanox TECHNOLOGIES
DELL
Tencent 腾讯
恒为科技 EmbedWay
DELTA
PEGATRON 和碩聯合科技
metaswitch
Edge-corE NETWORKS
ufiSpace
Microsoft
MARVELL
Inventec
CAVIUM
Nephos
Quanta Computer
WNC Wistron NeWeb Corp.
Innovium™

Newly Joined Members since 2019

SONiC – Warm Boot for High Reliability



Warm Boot: A True Community Effort

BROADCOM

Alibaba Group

apstra

Innovium™

BAREFOOT NETWORKS

CISCO

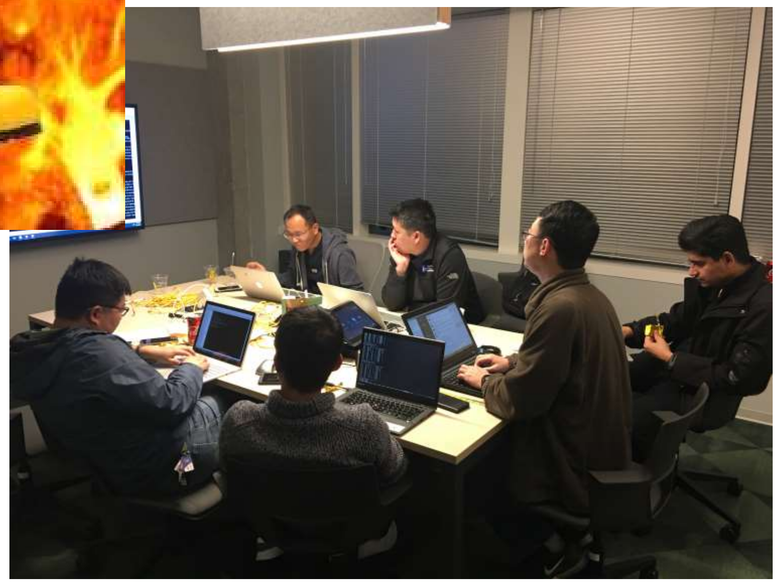
Microsoft

ARISTA

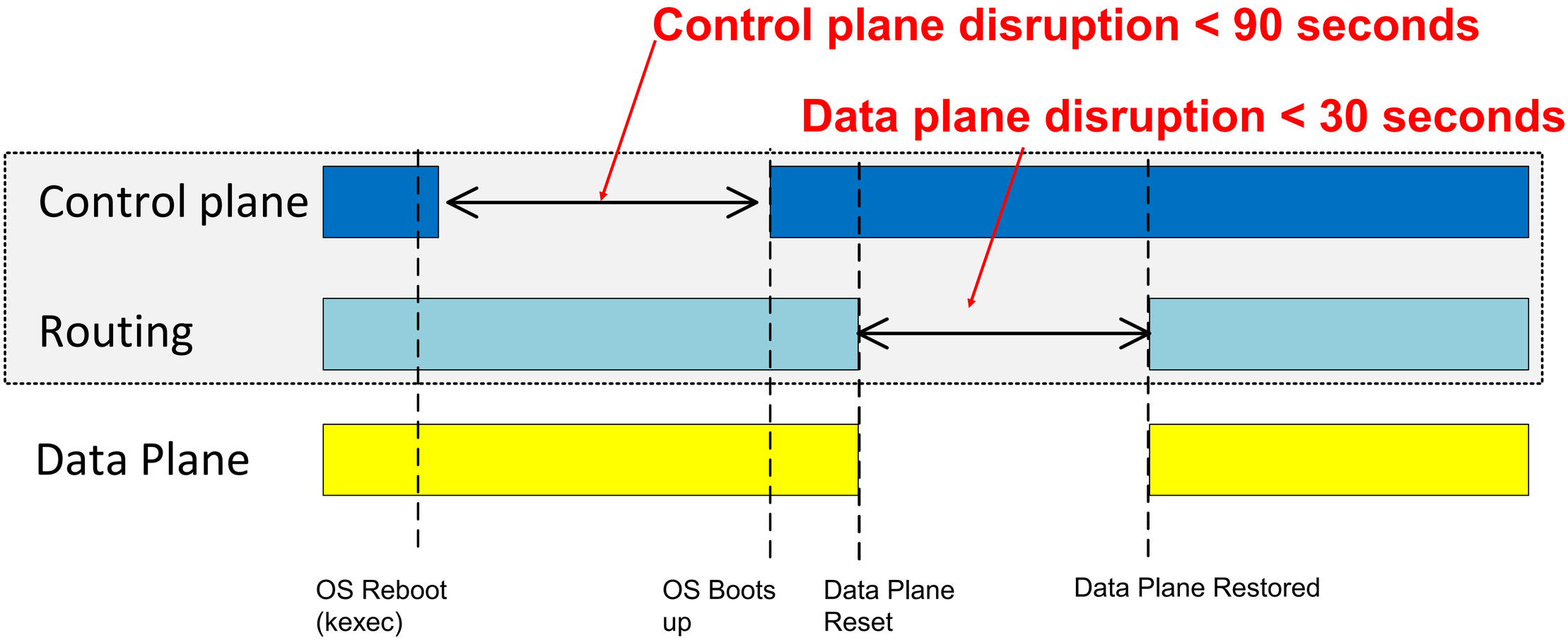
Mellanox TECHNOLOGIES

DELL

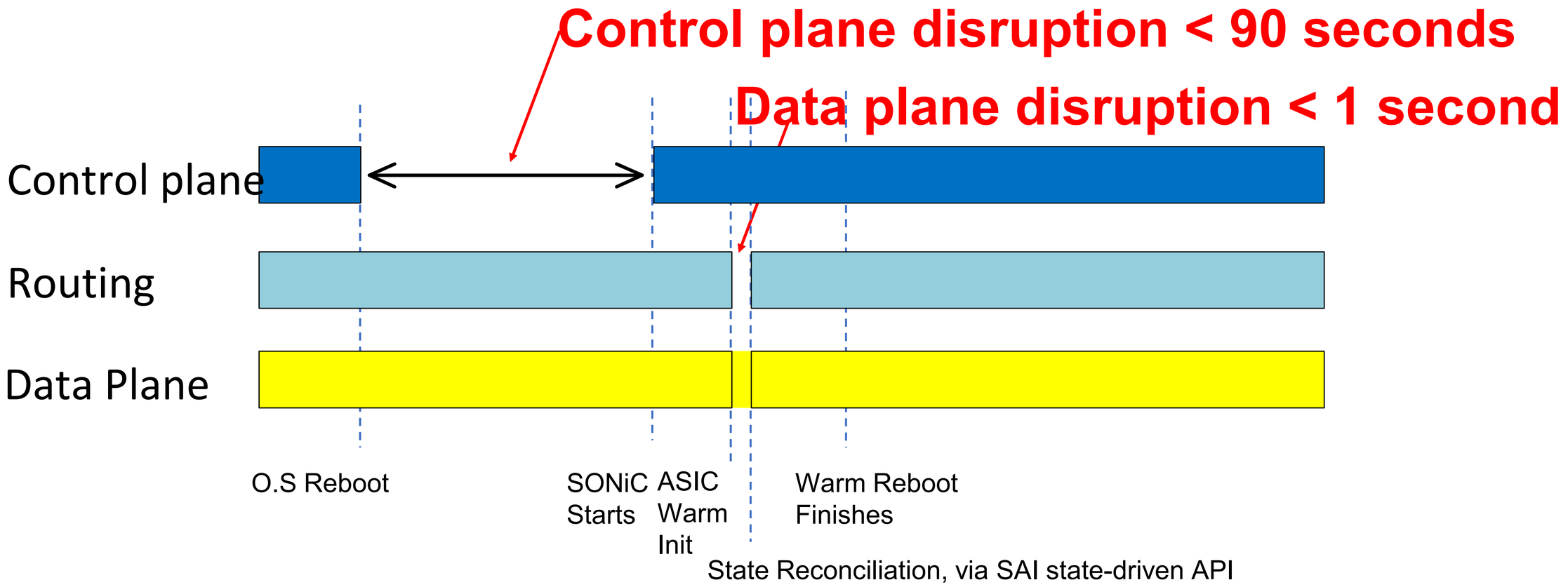
Nephos



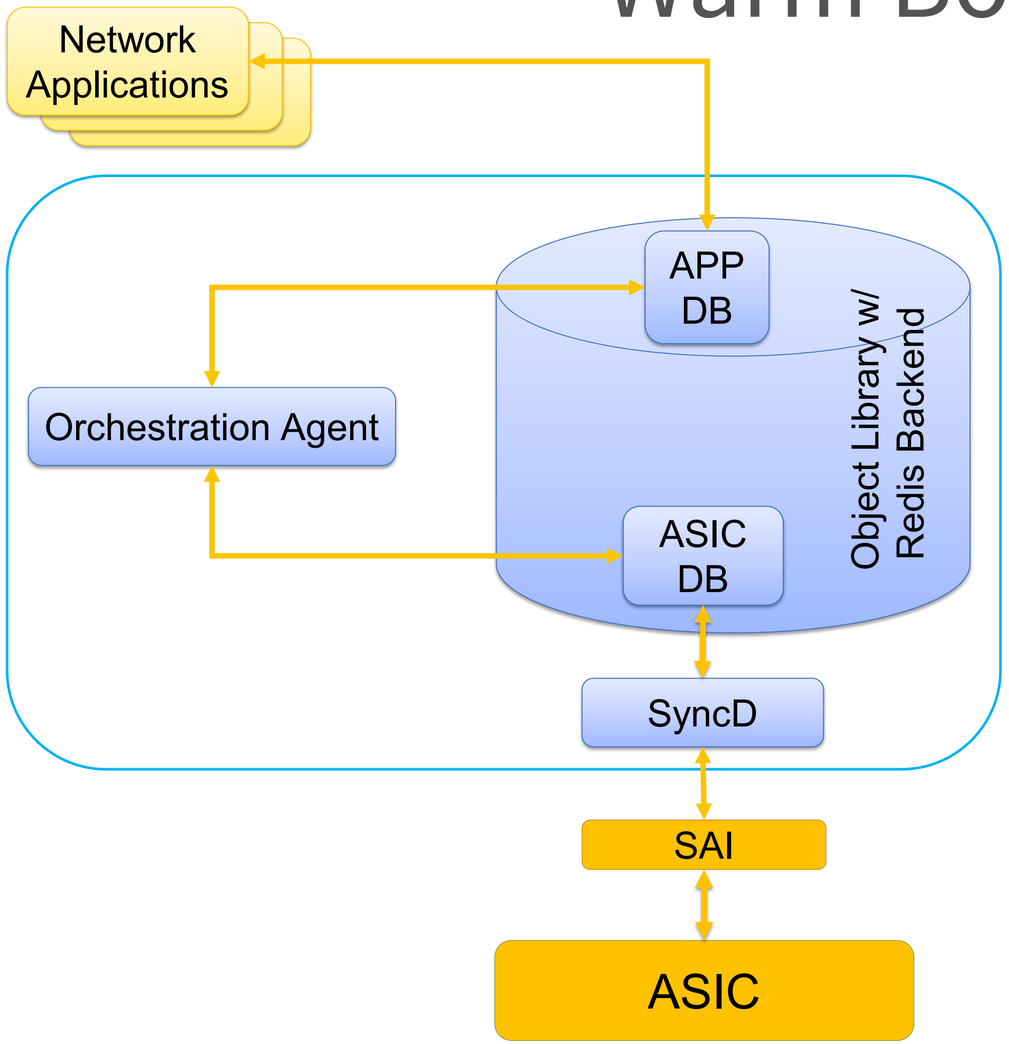
Fast Boot


Control plane disruption < 90 seconds
Data plane disruption < 30 seconds
Control plane
Routing
Data Plane
OS Reboot (kexec)
OS Boots up
Data Plane Reset
Data Plane Restored

Warm Boot

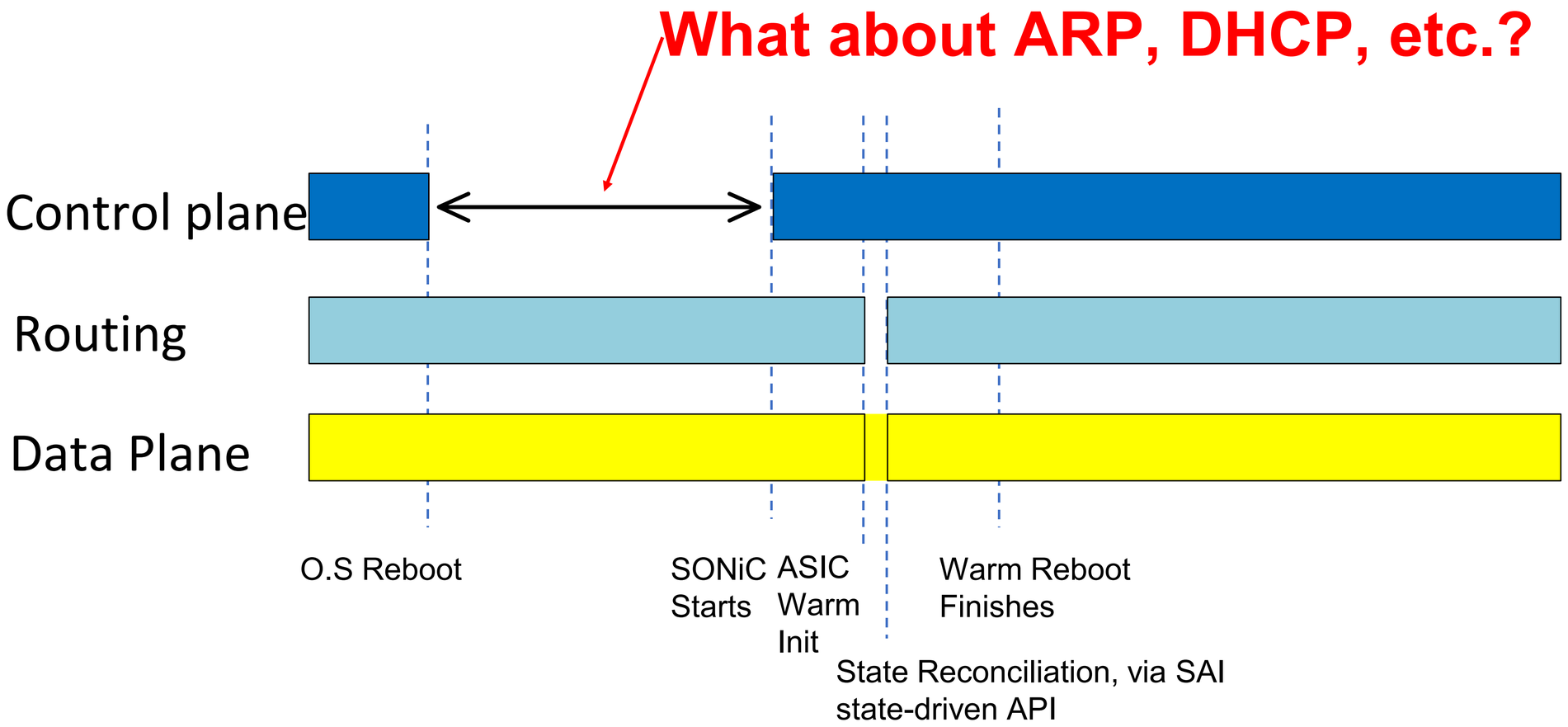

Control plane disruption < 90 seconds
Data plane disruption < 1 second
Control plane
Routing
Data Plane
O.S Reboot
SONiC Starts
ASIC Warm Init
Warm Reboot Finishes
State Reconciliation, via SAI state-driven API

Warm Boot Architecture

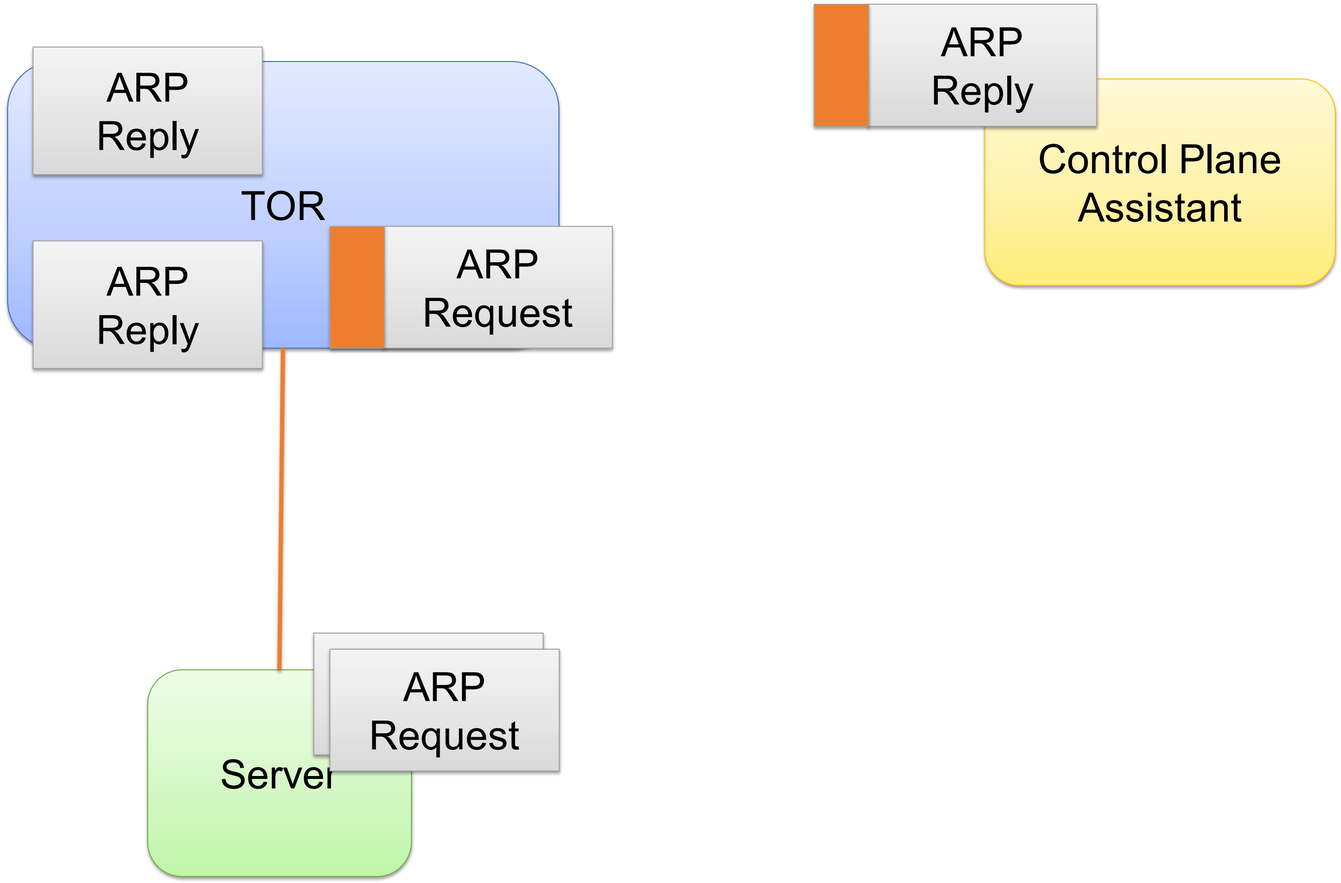


1. Warm boot script stores App/ASIC DB on disc
2. Redis restores App/ASIC DB after reboot
3. OA reads AppDB and compiles a new ASIC DB
4. SyncD compares old/new ASIC DB, and apply diff to the ASIC
5. Applications waking up in parallel
   - May staged changes to App DB
   - OA comes in as usual, updates ASIC dB
   - SyncD keeps syncing ASIC DB to hardware

We Are Not Done Yet – Control Plane?

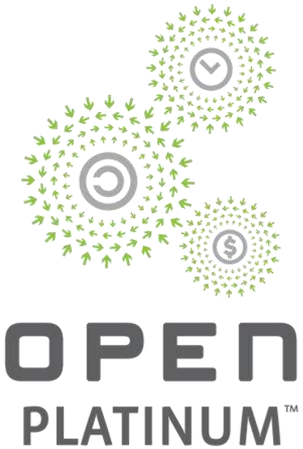


Control Plane Assistant



- TOR ASIC encaps ARP and send to CPA
- CPA responds with ARP reply
- TOR ASIC decaps the ARP and sends to the server

SONiC - Management Framework



Broadcom and Dell

Broadcom and SONiC

- Broadcom is contributing heavily to the success of the SONiC project
  - Open-source
  - Cloud, DC, Enterprise use-cases
- Engaging closely with Community (upstream, reviews, testing, support etc)

- Features contributed include: -
  - Now (201910)
    - ZTP, NAT, STP/PVST, BFD, L2/L3 Enhancements, MMU Thresholds, Platform Development Kit, VRF-Lite, Error Handling, Debug Framework, Core dump file handling, Build Improvements
    - Management Framework
  - Next
    - EVPN/VXLAN, M-LAG, IP Multicast (IGMP, PIM-SSM), VRRP, IGMP Snooping, MSTP, RADIUS, IPv6 Improvements, PTP, Instrumentation/Telemetry
    - Management Framework Improvements, FRR Management Integration, RBAC
- Items in red are a joint effort with Dell, and is our main topic today

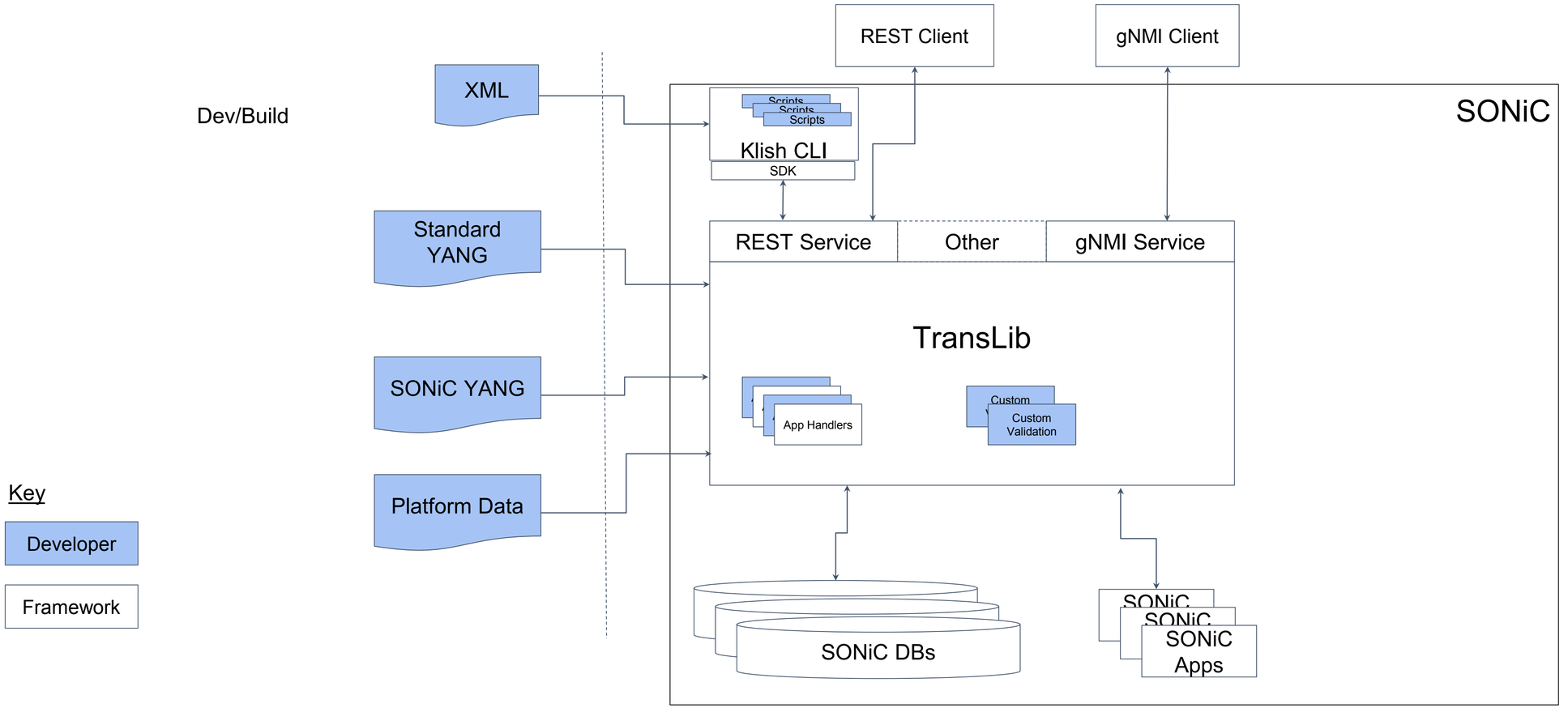
Management Framework Goals

- Integrated management experience for SONiC
  - Industry standard CLI
  - Standards-based Programmatic Interfaces (e.g. OpenConfig)
  - OEM-style AAA/Security
  - Broad Feature Coverage

- Create a Framework to allow: -
  - Rapid UI development from standard or custom data models
    - CLI, REST/RESTCONF, gNMI etc
  - Full configuration validation and error response

- Start the process of filling out UI content

The Big Picture

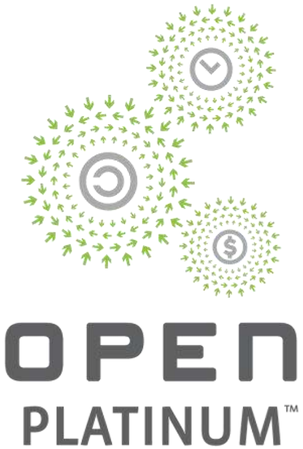


Implementation Work

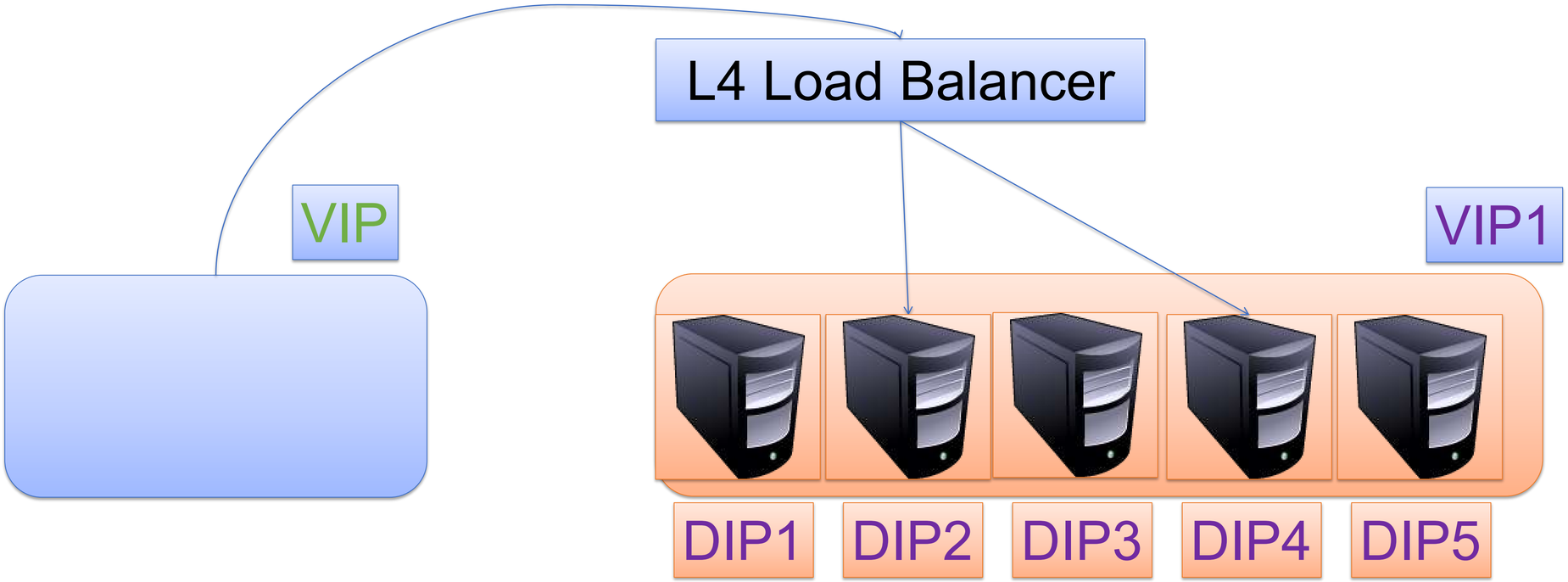
- SONiC 201910
  - Complete Framework, along with Developer Training and Guidelines
    - See https://github.com/Azure/SONiC/pull/436 for more
  - Defined guidelines for writing “SONiC YANG”
  - Feature implementations
    - gNMI/OpenConfig - interfaces, LLDP, System, Platform, ACL
    - Supporting IS-CLI for the above
    - REST service

- Future releases
  - Framework Improvements and Optimization
    - Taking feedback from the User Community
  - Full Privilege-level based Authorization, RBAC
  - Much more feature content (including SONiC legacy features)
    - Incl. FRR

SONiC – Extensibility For New Scenarios



Layer-4 Load Balancing



Layer-4 load balancing is a critical function

- handle both inbound and inter-service traffic
- >40%* of cloud traffic needs load balancing (Ananta [SIGCOMM'13])

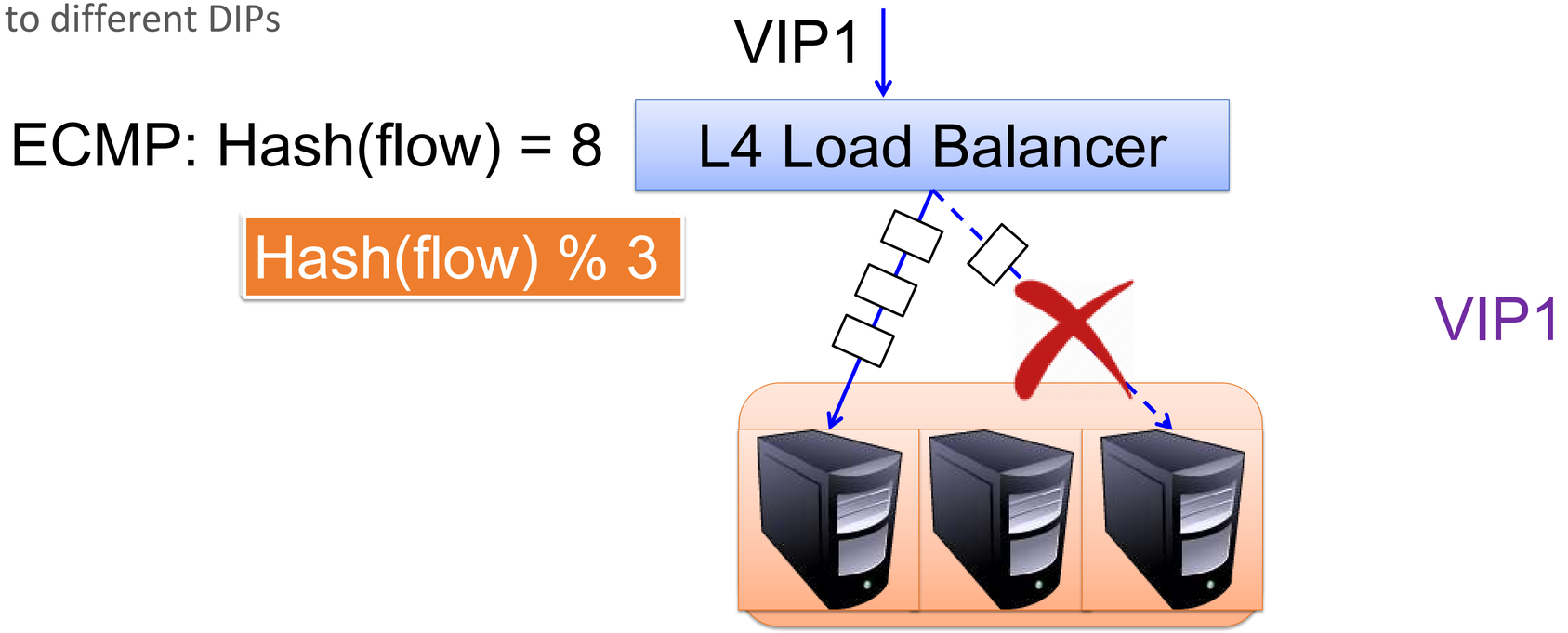
Frequent DIP pool updates

DIP pool updates

- failures, service expansion, service upgrade, etc.
- up to 100 updates per minute in a Big cluster

Hash function changes under DIP pool updates

- packets of a connection get to different DIPs
- connection is broken



Layer-4 Load Balancing

Per-connection consistency

- Broken connections degrade the performance of cloud services
  - tail latency, service level agreement, etc.

- PCC: all the packets of a connection go to the same DIP

L4 load balancing needs connection states

L4 load balancing needs connection states in HW?

Programmable ASICs allows such states in HW

Load balancer on SONiC

- SONiC Provide basic functions for managing switches
  - L2/L3 forwarding
  - Management plane such as LLDP, SNMP, Telemetry
- Extending SONiC
  - Introduce new config db entries
  - Extend SwSS to provide VIP-to-DIP management
  - Extend SAI to manage programmable ASICs

SONiC – Load Balancer Config DB Schema

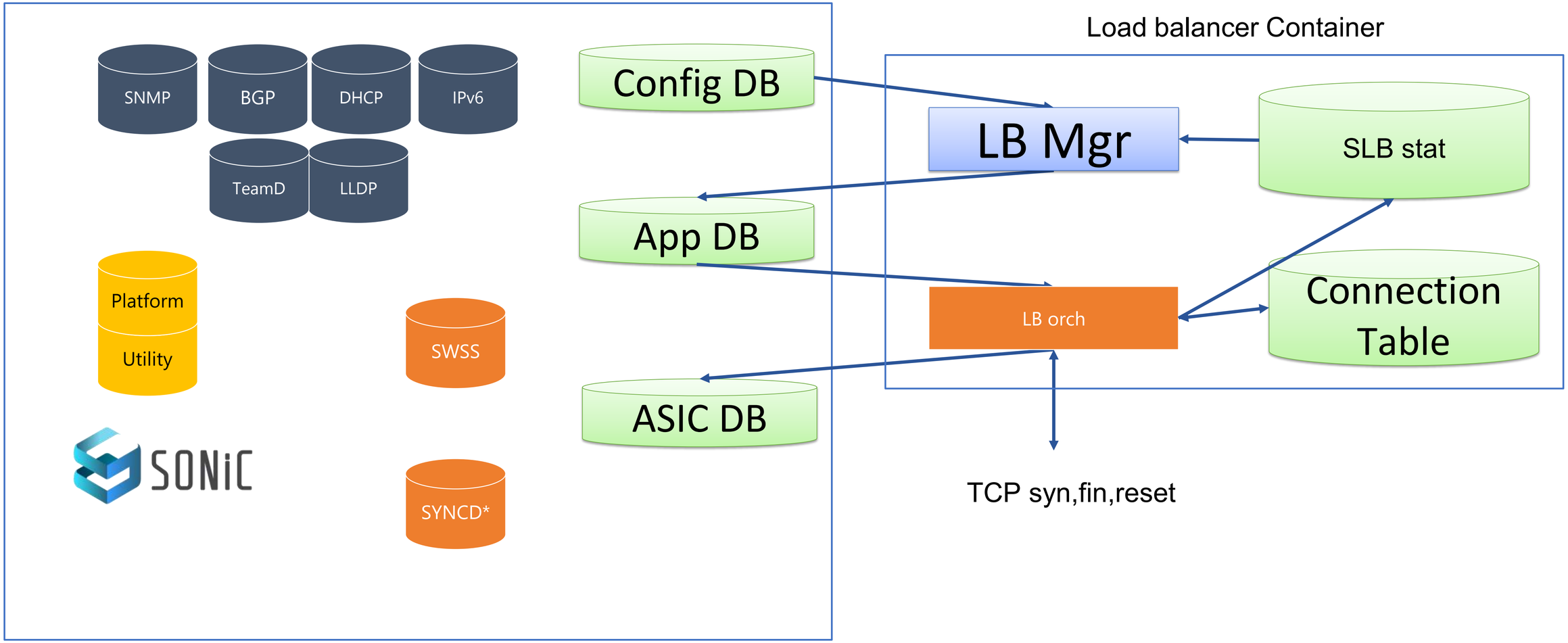
DIP table

- Key
  - DIP
- Data
  - Weight(optional)
  - state {active, disabled}
  - DMAC
  - underlay DIP
  - VNI

VIP table

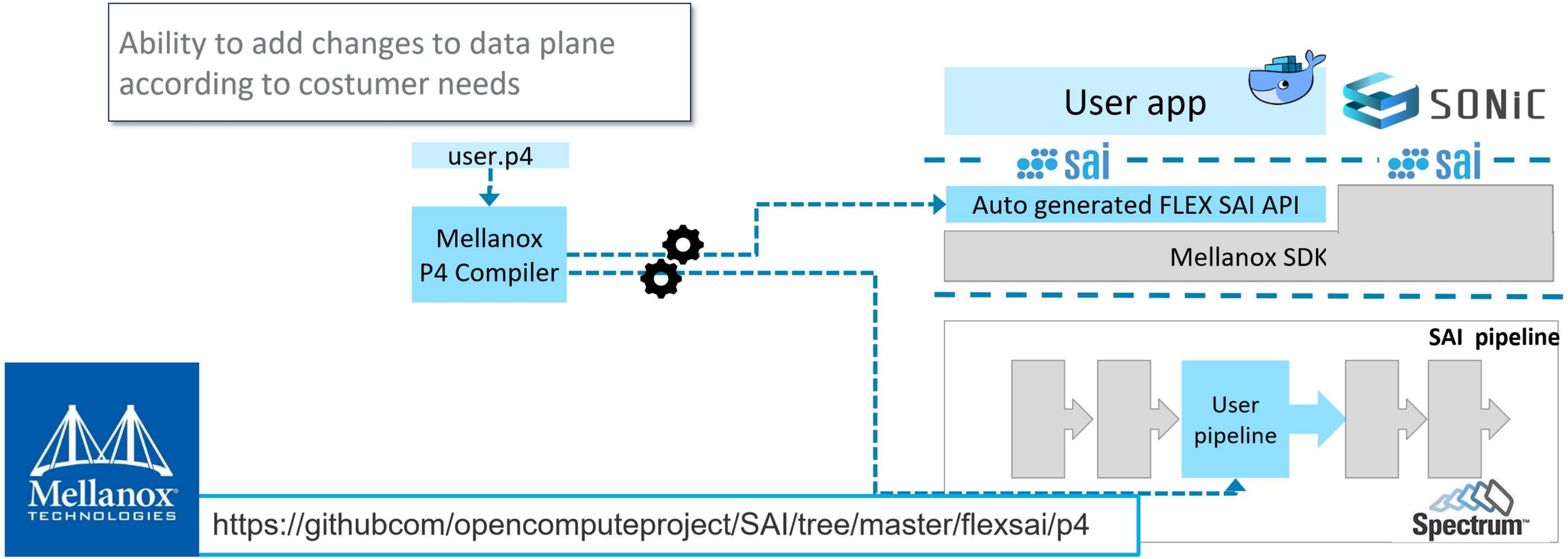
- Key
  - VIP
- Data
  - {list of DIP}
  - Num of DIP(optional)

Extending SwSS on SONiC



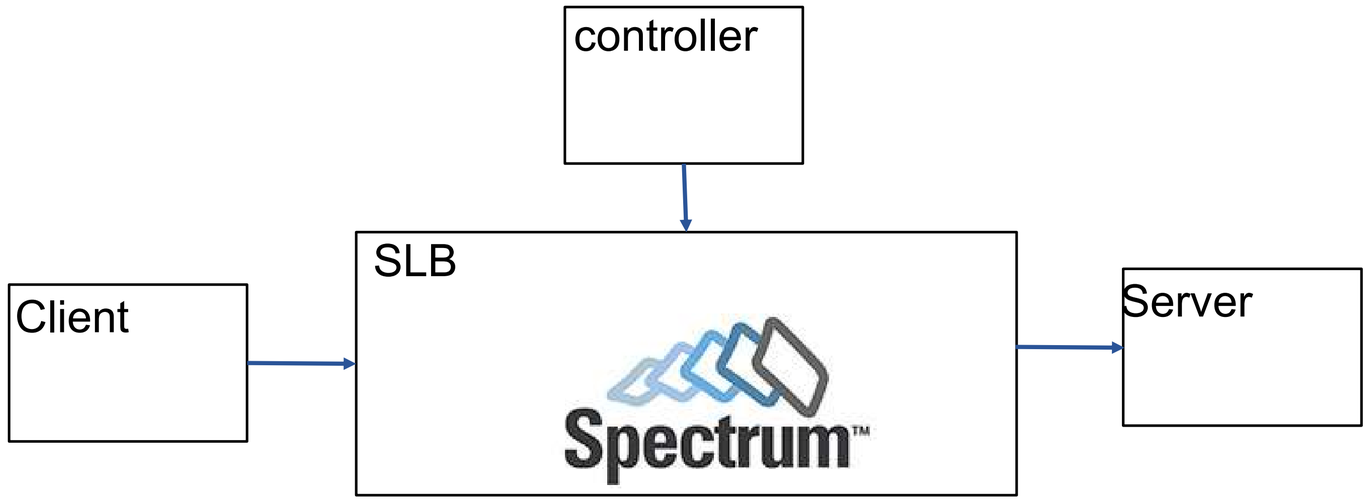
FlexSAI Extension

Ability to add changes to data plane according to costumer needs

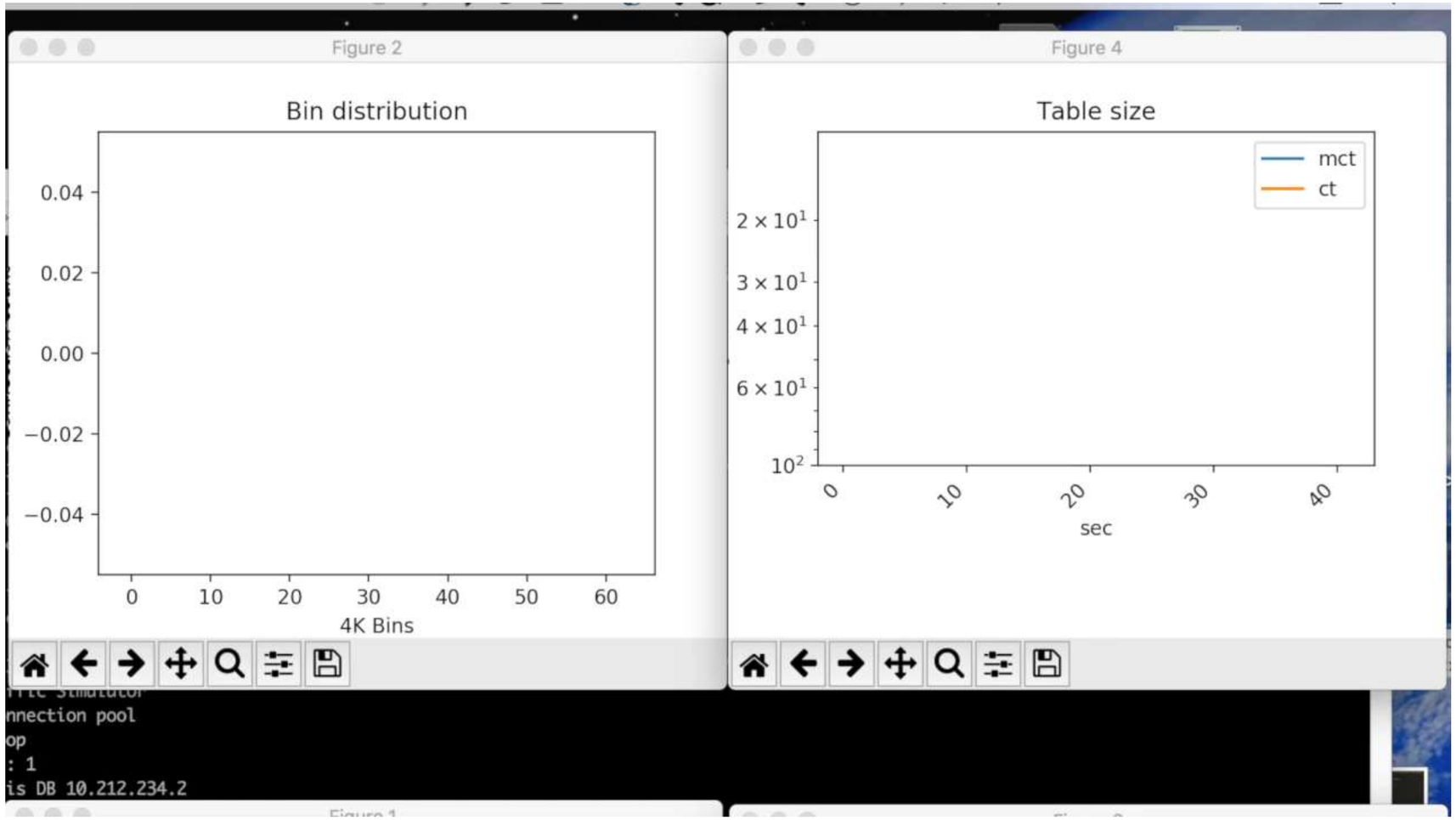


Demo

- Client
  - Generate 100K connection per sec
  - Average connection live time 10 sec
- SLB box
  - Load balancer
  - Single VIP
  - 1K DIP
- Controller
  - Create DIP change in average every 10 sec
- Server
  - Receive and monitor connection



Bin distribution and hardware CT size



Open Invitation

Inviting contributions in all areas

- SONiC/SAI
- Hardware platform
- New features, applications, tests and tools
- Download, test, Deploy!

Website: https://azure.github.io/SONiC/

Source code: https://github.com/Azure/SONiC/blob/gh-pages/sourcecode.md

We will be in Room G108 from 12:00 ~ 2:30pm today for Q&A, welcome !


OCP
REGIONAL
SUMMIT
Open. Together.
OCP Regional Summit
26–27, September, 2019