SCTE CABLE-TEC EXPO 22
SEPTEMBER 19-22 • PHILADELPHIA, PA

2022 Fall Technical Forum
SCTE • CABLELABS • NCTA

Peas In a Pod: What Makes Them Green?

A Technical Paper prepared for SCTE by

Defu Li
Distinguished Engineer
Comcast Cable
Massachusetts
+1 (267) 586-7680
Defu_Li@comcast.com

Richard Grivalsky
Senior Energy Engineer
Comcast Cable
+1 (802) 316-6553
Richard_Grivalsky@cable.comcast.com

Robert Gaydos, Comcast Cable

Ashok Ramasamy, Comcast Cable

Eric Stonfer, Comcast Cable

Gianni DiGregorio, Comcast Cable

Table of Contents

Title **Page Number**

1. Introduction ... 3
1.1. vCMTS and DAA ... 3
1.1. vBNG ... 3
1.2. PPODs ... 3
1.3. Carbon Neutrality and Purpose ... 4
2. Observability ... 4
2.1. Framework ... 4
2.2. Collector and Monitor ... 5
2.2.1. Collectd ... 5
2.2.2. IPMI Plugin ... 6
2.2.3. Turbostat Plugin ... 6
2.2.4. Grafana Dashboard ... 7
3. Hot Standby & Power Saving Mode ... 8
3.1. Hot Standby ... 8
3.2. Intel CPU Power Saving Mode ... 11
3.3. Future Considerations ... 12
4. Conclusion ... 14

Abbreviations ... 14

Bibliography & References ... 15

List of Figures

Title **Page Number**

Figure 1 - MHAv2 [Source: CM-SP-R-PHY Specification] ... 3
Figure 2 - Logical View of Access Network and Compute Nodes in a Leaf/Spine Architecture ... 4
Figure 3 - PPOD Observability Framework ... 5
Figure 4 - Snippets of Prometheus Configmap and Collectd Pod Spec[1] ... 6
Figure 5 - IPMI Metrics ... 6
Figure 6 - Screenshot of Turbostats Output[1] ... 7
Figure 7 - Power Consumption Dashboard[1] ... 7
Figure 8 - Dual-Redundant vs. Hot Standby Line Drawing ... 8
Figure 9 – Server Rear Elevation Dual-Redundant vs. Hot Standby Watts Consumed ... 10
Figure 10 - Power Consumption for Host with C-State Disabled[1] ... 11
Figure 11 - Power Consumption of a Host with C-State Enabled[1] ... 12
Figure 12 - Workload Utilization Percentage for a PPOD for One Week[1] ... 13
Figure 13 - Stacked Workload Utilization Percentage for a PPOD for One Week[1] ... 13

List of Tables

Title **Page Number**

Table 1 - Hot Standby Enabled Then Disabled Data Segment ... 9
Table 2 - B-Side Breakers Closed ... 10
Table 3 - B-Side Breakers Open ... 10

1. Introduction

1.1. vCMTS and DAA

The distributed access architecture (DAA) specification, or modular head-end architecture version 2 (MHAv2), was introduced to address cable headend space and power limitations. The traditional integrated CMTS (iCMTS) or cable converged access platform (CCAP) functions were split into two: the physical (PHY) function, and the core function. The remote PHY device (RPD) provides the PHY function, while the core functions consist of CMTS and CCAP operating on the MAC or IP layers.

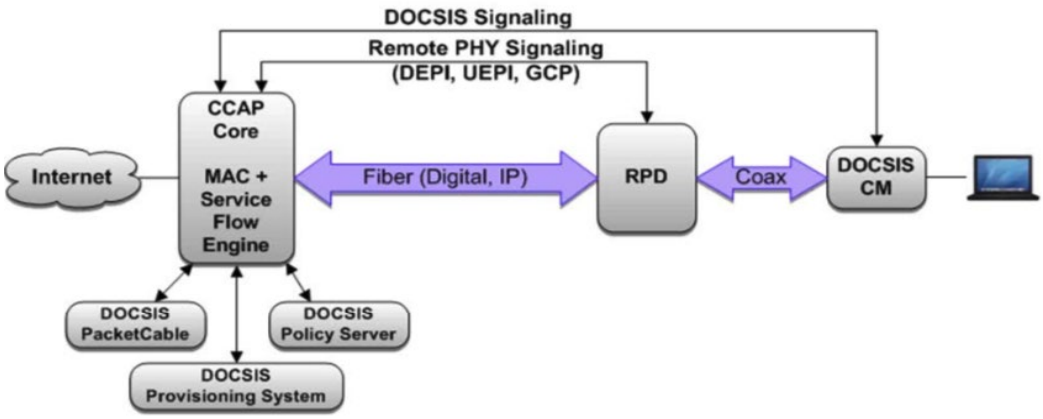


Figure 1 - MHAv2 [Source: CM-SP-R-PHY Specification]

The split allows the core functions to run on a cloud computing platform. The virtualized CMTS/CCAP (vCMTS/vCCAP) is a collection of software applications, built upon the microservice architecture pattern and targeted for cloud computing platforms. Comcast Cable has built its own private cloud in order to host these vCMTS software applications.

1.1. vBNG

CableLabs' DOCSIS Provisioning of EPON (DPoE) specification enables an operator to deploy EPON technology using the existing DOCSIS based backend systems. This specification allows an optical network unit (ONU) to be emulated as virtual cable modem (vCM).

The Comcast Private Cloud can host the virtual broadband gateway (vBNG) application which supports EPON technology. Like DPoE, vBNG emulates an ONU in order to utilize DOCSIS based network device provisioning backend systems.

1.2. PPODs

A typical private cloud is likely to consist of many server racks. Like peas in a pod (PPOD), these server racks are built identically providing operational efficiency and easy scalability. There are thousands of 'peas' (servers) spread across hundreds of PODs in an operator's network.

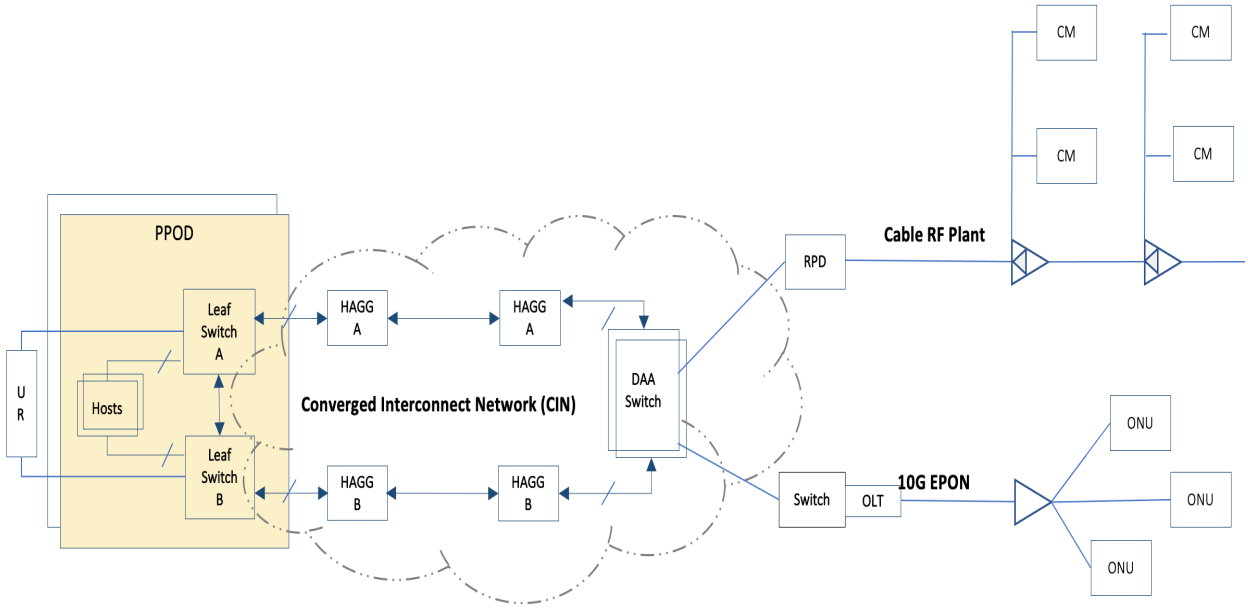


Figure 2 - Logical View of Access Network and Compute Nodes in a Leaf/Spine Architecture

The PPODs are deployed across hundreds of sites nationwide. Each PPOD contains several servers which form a compute cluster, each server has dual ethernet ports connecting to a pair of leaf switches. vCMTS and vBNG workloads are deployed and replicated on these PPODs OLT and RPD traffic is tunneled to or from the vBNG and vCMTS via the converged interconnect network (CIN) via a leaf-spine switch fabric. Upstream traffic to or from the internet is routed via upstream routers (URs)

1.3. Carbon Neutrality and Purpose

Looking ahead to Comcast's commitment to being carbon neutral by 2035, the question becomes, what can we do to "green-up" our PPODs, make them more energy efficient, and in the process reduce our operational expenditures?

In this paper, we will discuss what is involved to provide energy consumption observability for the Comcast Private Cloud. We will discuss the techniques that we employ to provide immediate energy saving as well as more advance techniques based on load and demand characteristics of our containerized network function (CNF) workloads. This paper will conclude with the lessons learned and future strategy for energy efficiency looking beyond the PPODs, in the wider Comcast ecosystem.

2. Observability

2.1. Framework

Since the start real time observability has been crucial for the Comcast Cable Private Cloud, as such we have built a stack based upon on Elasticsearch, Logstash, and Kibana (ELK) stack and Prometheus, a time series database (TSDB).

At a high level, the Prometheus server periodically scrapes metrics from targets in the PPOD. The long-term metrics are pushed to S3 ThanosStore, with Grafana dashboards providing a human friendly interface. ThanosQuery provides a distributed query engine for short-term locally cached metrics.

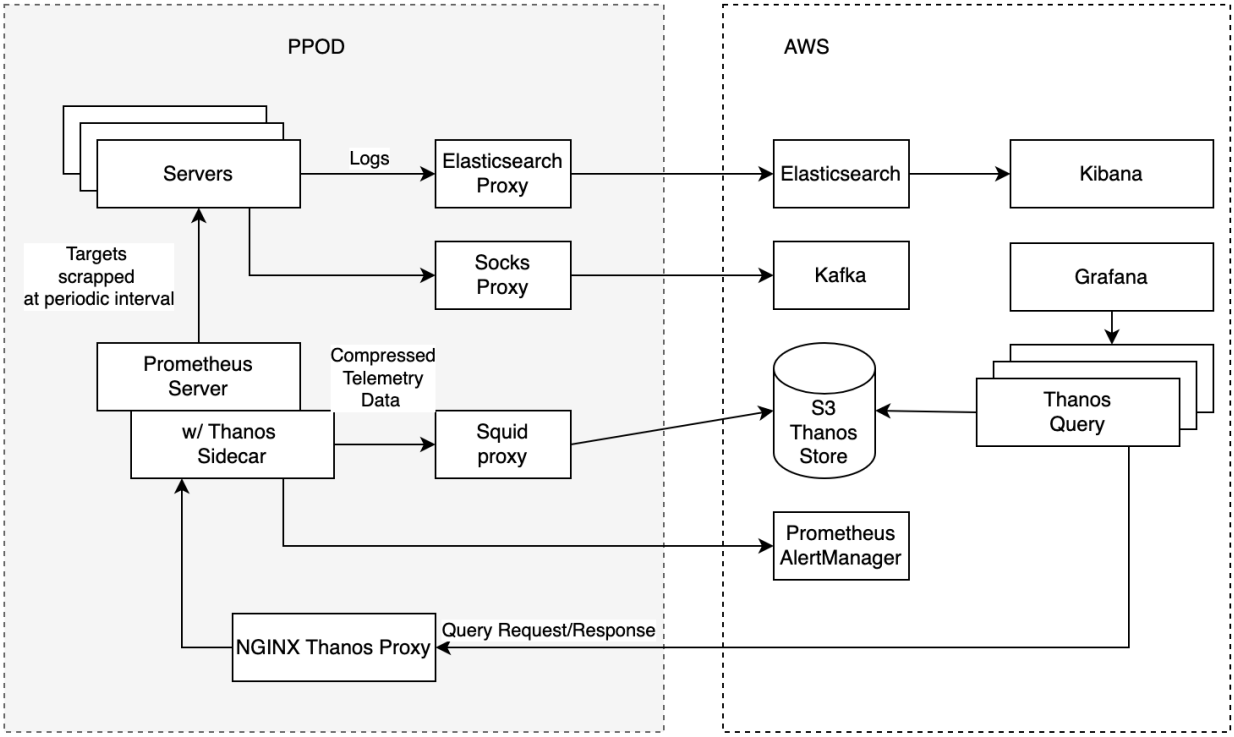


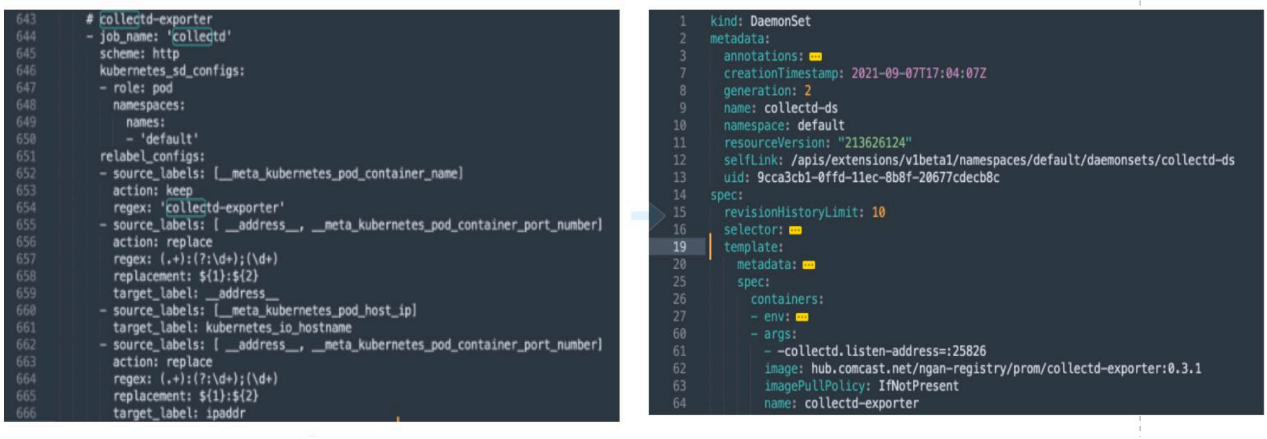
Figure 3 - PPOD Observability Framework

In the subsections which follow, we describe the components which provide power consumption metrics.

2.2. Collector and Monitor

2.2.1. Collectd

Collectd is a Linux daemon that collects, stores and transfers performance metrics on a per host level. Collectd is deployed as a DaemonSet for all hosts, in each PPOD. Prometheus in turn scrapes the metrics provided by Collectd.



```
# collectd-exporter
- job_name: 'collectd'
  scheme: http
  kubernetes_sd_configs:
  - role: pod
    namespaces:
      names:
      - 'default'
  relabel_configs:
  - source_labels: [__meta_kubernetes_pod_container_name]
    action: keep
    regex: 'collectd-exporter'
  - source_labels: [ __address__, __meta_kubernetes_pod_container_port_number]
    action: replace
    regex: (.+):(?:\d+);(\d+)
    replacement: ${1}:${2}
    target_label: __address__
  - source_labels: [__meta_kubernetes_pod_host_ip]
    target_label: kubernetes_io_hostname
  - source_labels: [ __address__, __meta_kubernetes_pod_container_port_number]
    action: replace
    regex: (.+):(?:\d+);(\d+)
    replacement: ${1}:${2}
    target_label: ipaddr
```

```
kind: DaemonSet
metadata:
  annotations: ...
  creationTimestamp: 2021-09-07T17:04:07Z
  generation: 2
  name: collectd-ds
  namespace: default
  resourceVersion: "213626124"
  selfLink: /apis/extensions/v1beta1/namespaces/default/daemonsets/collectd-ds
  uid: 9cca3cb1-0ffd-11ec-8b8f-20677cdecb8c
spec:
  revisionHistoryLimit: 10
  selector: ...
  template:
    metadata: ...
    spec:
      containers:
      - env: ...
      - args:
        - -collectd.listen-address=:25826
        image: hub.comcast.net/ngan-registry/prom/collectd-exporter:0.3.1
        imagePullPolicy: IfNotPresent
        name: collectd-exporter
```

Figure 4 - Snippets of Prometheus Configmap and Collectd Pod Spec[1]

2.2.2. IPMI Plugin

Collectd supports numerous loadable plugins. The Intelligent Platform Management Interface (IPMI) plugin uses the OpenIPMI library to read hardware sensors on the host in order to provide power consumption metrics as shown in the figure below.

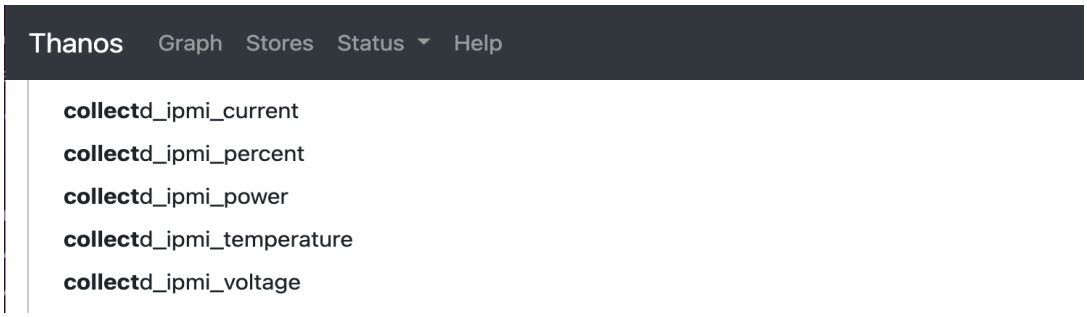


Figure 5 - IPMI Metrics

2.2.3. Turbostat Plugin

Turbostat is a Linux tool that reports processor frequency and statistics. The Turbostats Plugin utilizes Turbostats for reporting processor performance metrics.

Package t	Core	CPU RAMWatt	PKG_% RAM_%	Avg_MHz	Busy%	Bzy_MHz	TSC_MHz	IRQ	SMI	POLL	C1	C1E	C6	POLL%	C1%	C1E%	C6%	CPU%c1	CPU%c6	CoreTmp	PkgTmp	PkgWat
-	-	-	63	3.02	2100	2094	108432	0	2280	123551	0	0	0.00	96.95	0.00	0.00	96.98	0.00	53	53	61.650
.00	0.00	0.00																			
0	0	0	69	3.30	2100	2095	2969	0	177	3622	0	0	0.01	96.74	0.00	0.00	96.70	0.00	50	52	32.100
.00	0.00	0.00																			
0	0	16	53	2.55	2100	2095	3551	0	20	4062	0	0	0.00	97.48	0.00	0.00	97.45				
0	1	1	67	3.18	2100	2095	3286	0	253	3972	0	0	0.01	96.86	0.00	0.00	96.82	0.00	49		
0	1	17	59	2.82	2100	2095	3983	0	3	4377	0	0	0.00	97.23	0.00	0.00	97.18				
0	2	2	43	2.04	2100	2095	3330	0	12	3583	0	0	0.00	98.00	0.00	0.00	97.96	0.00	49		
0	2	18	146	6.96	2100	2095	4254	0	7	4615	0	0	0.00	93.08	0.00	0.00	93.04				
0	3	3	75	3.58	2100	2095	3167	0	17	3538	0	0	0.00	96.46	0.00	0.00	96.42	0.00	49		
0	3	19	56	2.69	2100	2095	5258	0	883	5577	0	0	0.03	97.36	0.00	0.00	97.31				
0	4	4	76	3.62	2100	2095	3110	0	2	3456	0	0	0.00	96.42	0.00	0.00	96.38	0.00	50		
0	4	20	66	3.15	2100	2095	4401	0	236	4802	0	0	0.01	96.90	0.00	0.00	96.85				
0	5	5	153	7.29	2100	2095	2910	0	3	3236	0	0	0.00	92.75	0.00	0.00	92.71	0.00	49		
0	5	21	65	3.09	2100	2095	2869	0	24	3445	0	0	0.00	96.94	0.00	0.00	96.91				
0	6	6	42	2.01	2100	2095	3797	0	119	4170	0	0	0.01	98.04	0.00	0.00	97.99	0.00	49		
0	6	22	49	2.32	2100	2095	3112	0	213	3944	0	0	0.01	97.72	0.00	0.00	97.68				
0	7	7	98	4.67	2100	2095	3233	0	118	3634	0	0	0.01	95.37	0.00	0.00	95.33	0.00	50		
0	7	23	67	3.21	2100	2095	3271	0	9	3838	0	0	0.00	96.83	0.00	0.00	96.79				
1	0	8	45	2.16	2100	2095	2884	0	9	3151	0	0	0.00	97.88	0.00	0.00	97.84	0.00	52	53	29.620
.00	0.00	0.00																			
1	0	24	78	3.72	2100	2095	3619	0	13	3754	0	0	0.00	96.32	0.00	0.00	96.28				
1	1	9	58	2.79	2100	2095	3326	0	27	3740	0	0	0.00	97.25	0.00	0.00	97.21	0.00	52		
1	1	25	74	3.52	2100	2095	2801	0	3	3271	0	0	0.00	96.51	0.00	0.00	96.48				
1	2	10	54	2.56	2100	2095	2889	0	16	3980	0	0	0.00	97.48	0.00	0.00	97.44	0.00	52		
1	2	26	46	2.18	2100	2095	3611	0	9	3887	0	0	0.00	97.86	0.00	0.00	97.82				
1	3	11	67	3.21	2100	2095	4087	0	13	4848	0	0	0.00	96.84	0.00	0.00	96.79	0.00	51		
1	3	27	39	1.87	2100	2095	2992	0	0	3317	0	0	0.00	98.16	0.00	0.00	98.13				
1	4	12	46	2.19	2100	2095	3071	0	3	3543	0	0	0.00	97.85	0.00	0.00	97.81	0.00	52		
1	4	28	45	2.13	2100	2095	2817	0	0	3140	0	0	0.00	97.90	0.00	0.00	97.87				
1	5	13	43	2.03	2100	2095	3005	0	66	4185	0	0	0.01	98.01	0.00	0.00	97.97	0.00	53		
1	5	29	50	2.37	2100	2095	3043	0	10	3774	0	0	0.00	97.67	0.00	0.00	97.63				
1	6	14	42	1.99	2100	2095	3941	0	11	4276	0	0	0.00	98.06	0.00	0.00	98.01	0.00	51		
1	6	30	44	2.12	2100	2095	2588	0	0	3553	0	0	0.00	97.92	0.00	0.00	97.88				
1	7	15	44	2.08	2100	2095	3168	0	1	3434	0	0	0.00	97.96	0.00	0.00	97.92	0.00	52		
1	7	31	71	3.40	2100	2095	3189	0	3	3827	0	0	0.00	96.64	0.00	0.00	96.60				
Package t	Core	CPU RAMWatt	Avg_MHz RAM_%	Busy%	Bzy_MHz	TSC_MHz	IRQ	SMI	POLL	C1	C1E	C6	POLL%	C1%	C1E%	C6%	CPU%c1	CPU%c6	CoreTmp	PkgTmp	PkgWat

Figure 6 - Screenshot of Turbostats Output[1]

2.2.4. Grafana Dashboard

Power consumption data visualization can be easily created and customized via a Grafana Dashboard. The figure below shows the power consumption by host and by PPOD. The charts on the left show the total power consumption by host and by PPOD. The charts on the right show the breakdown by the power supply unit by host and by PPOD.

Figure 7 - Power Consumption Dashboard[1]

3. Hot Standby & Power Saving Mode

3.1. Hot Standby

Power supply unit (PSU) hot standby, also referred to as hot sparing, is the ability for a single power supply to transform input to platform required voltage while keeping an idle power supply in reserve, as seen in Figure 8. The platform stages PSUs on or off dependent upon the platform's throughput and required power. The power supply in standby configuration does not transform input voltage to platform required voltages but does maintain telemetry, connection to the common buss, and is instantaneously available to support higher energy demand or in support of loss to the active-primary power supply.

It is through this idle state operators can realize a reduction in energy consumption. In dual redundant mode, the platform's required load will be split ≈50% on each PSU. If the PSU has an 800-watt capacity and is only loaded with 100-watts (12.5%), it may not be optimally loaded. This creates transformation through two PSUs and impacts energy dissipated to transform voltage; total dissipation is dependent on the efficiency curve of the PSU and the output load. If we optimize the load, by enabling hot standby, we can improve the efficiency and reduce the number of locations voltage is transformed. By moving from dual-redundant operation at 100-watts (12.5%) of PSU0 and PSU1's capacity utilized, hot standby operation loads PSU0 at 200-watts (25% capacity).

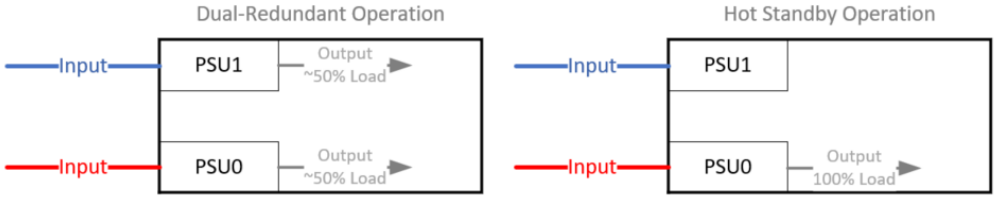


Figure 8 - Dual-Redundant vs. Hot Standby Line Drawing

Several original equipment manufacturers (OEMs) offer this platform setting today. In initial trials, a ~4.5% reduction in energy consumption for the deployed solution was documented.

Load-side power distribution measurements were captured with dual-redundant power supplies active for a minimum of (7) days prior to enabling of hot standby for the platforms tested as seen in Table 1. The platforms were then set to hot standby mode and measurements captured over time (30 days) before returning to dual-redundant mode for additional measurement.

During the hot standby trial time two distinct events were observed in which platform power exceeded the threshold for a single supply, the hot standby was brought into operational state, and then returned to hot standby with no service impact observed.

Table 1 - Hot Standby Enabled Then Disabled Data Segment

Date	Bus A Volt	Bus B Volt	Circ A03 Current	Circ B03 Current	Circuit 3 Total Watts	Circ A04 Current	Circ B04 Current	Circuit 4 Total Watts
11/28/2021 20:00	53.5	53.64341	8.389999	0.21	460.1300626	7.94	0.2	435.518682
11/29/2021 0:00	53.5	53.63208	8.41	0.21	461.1977368	7.98	0.2	437.656416
11/29/2021 4:00	53.5	53.654583	8.26	0.21	453.1774624	7.91	0.21	434.4524624
11/29/2021 8:00	53.5	53.665543	8.21	0.21	450.504764	7.86	0.21	431.779764
11/29/2021 12:00	53.5	53.650021	8.25	0.21	452.6415044	7.86	0.21	431.7765044
11/29/2021 20:00	53.599998	53.500065	4.81	4.12	478.2362582	5.57	2.89	453.1671767
11/30/2021 0:00	53.599998	53.500057	4.78	4.14	477.6982264	5.59	2.94	456.9141564
11/30/2021 4:00	53.599998	53.50074	4.76	4.12	475.5590393	5.57	2.93	455.3091571

For platforms which do not currently offer hot standby configuration, measurements were captured in dual-redundant state by aggregating channel loads as seen in Table 2. Channel 12, for example, was drawing a total of two amperes in dual-redundant mode. Power was removed from the B-side load distribution by way of opening breakers in order to test the specific platform load and its effect on PSU efficiency curve as seen in Table 3. Channel 12 was now only drawing 1.7 amperes. Energy avoidance was calculated with the nominal consumption of a PSU in hot standby factored for those devices by adding 13-watts of load, as seen in Figure 9. This process was performed on two separate PPODs of varying compute load and tested for 24-72 hours before restoring B-Side power.

Table 2 - B-Side Breakers Closed

Module 2A					
SNMP Chan Mapping	**Channel**	**Load**	**Ampacity**	**Inventory**	**Brkr Status**
22	12	1.2A	20A	YES	ON
23	13	2.1A	20A	YES	ON
24	14	2.2A	30A	YES	ON
25	15	2.3A	30A	YES	ON
Module 2B					
SNMP Chan Mapping	**Channel**	**Load**	**Ampacity**	**Inventory**	**Brkr Status**
32	12	0.8A	20A	YES	ON
33	13	2.2A	30A	YES	ON
34	14	1.8A	30A	YES	ON
35	15	2.2A	30A	YES	ON

Table 3 - B-Side Breakers Open

Module 2A					
SNMP Chan Mapping	**Channel**	**Load**	**Ampacity**	**Inventory**	**Brkr Status**
22	12	1.7A	20A	YES	ON
23	13	4.0A	20A	YES	ON
24	14	3.7A	30A	YES	ON
25	15	4.3A	30A	YES	ON
Module 2B					
SNMP Chan Mapping	**Channel**	**Load**	**Ampacity**	**Inventory**	**Brkr Status**
32	12	0.0A	20A	YES	OFF
33	13	0.0A	30A	YES	OFF
34	14	0.0A	30A	YES	OFF
35	15	0.0A	30A	YES	OFF

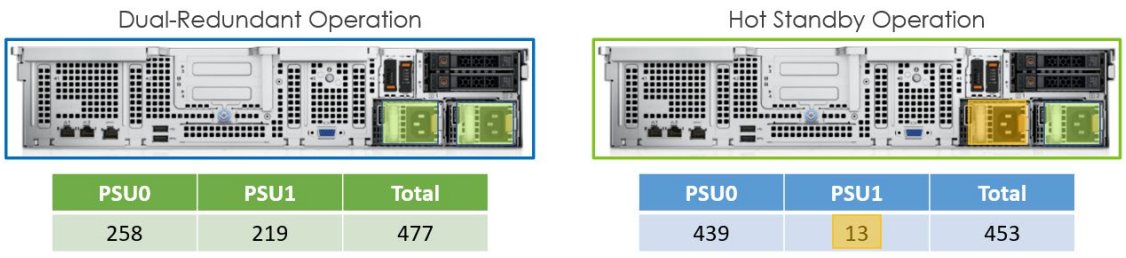


Figure 9 – Server Rear Elevation Dual-Redundant vs. Hot Standby Watts Consumed

As DAA expands the proliferation of server-based rack architectures, hot standby presents itself as a low-impact, reliable, and sustainable practice to aid in the drive toward carbon neutrality. Given the change and impact to installation and operational practices, processes must be built to ensure load is equally distributed across AC & DC plant circuits.

3.2. Intel CPU Power Saving Mode

All servers in a PPOD are Intel CPU based. Intel processors can be controlled by the following:

- Per core C-State
- Per core P-State

The C-State is an idle power state in which the processor is not executing instruction. The P-State is for various voltage or frequency levels in which the processor is still executing instructions.

For configuring and controlling C-states, on most modern Linux platforms C-states are automatically enabled, this is done via a combination of basic input/output system (BIOS) settings and the intel_idle driver. In order to dynamically force the system to a lower C-state (more power intensive) one can open the file /dev/cpu_dma_latency, and write a low value, usually (5) or under to this file. The value found in /dev/cpu_dma_latency represents the amount of latency in microseconds allowed for C-state transitions, by forcing this to a low value this should limit the CPU to C0 during active workloads and C1 during idle. For as long as this file remains open the C-states will be forced to these lower states.

Our first step was to understand C-state, workloads, and how they influence the power consumption of a host. This is accomplished by scheduling the workloads on a host, then measuring the power consumption data with C-state enabled and disabled.

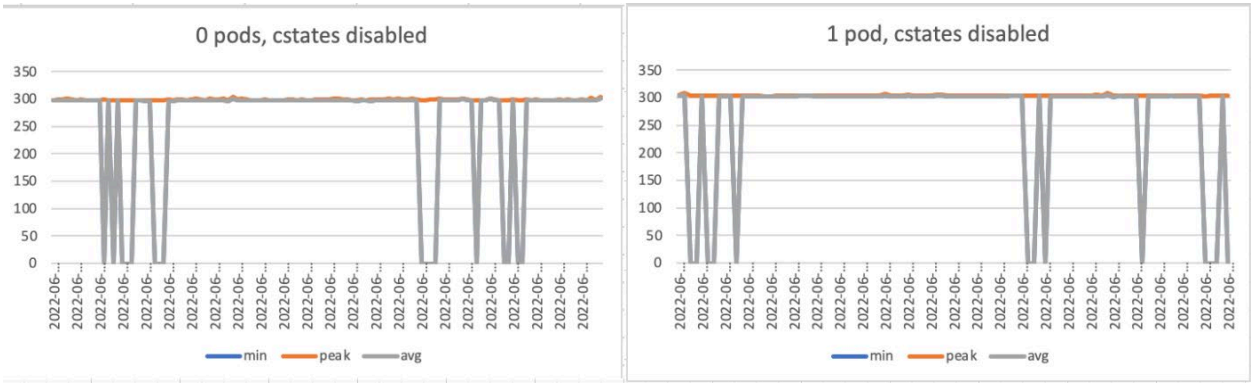


Figure 10 - Power Consumption for Host with C-State Disabled[1]

Our test result shows that with C-state saving mode disabled, the power consumption level remains constant for any number of CNF workloads scheduled on to a given host.

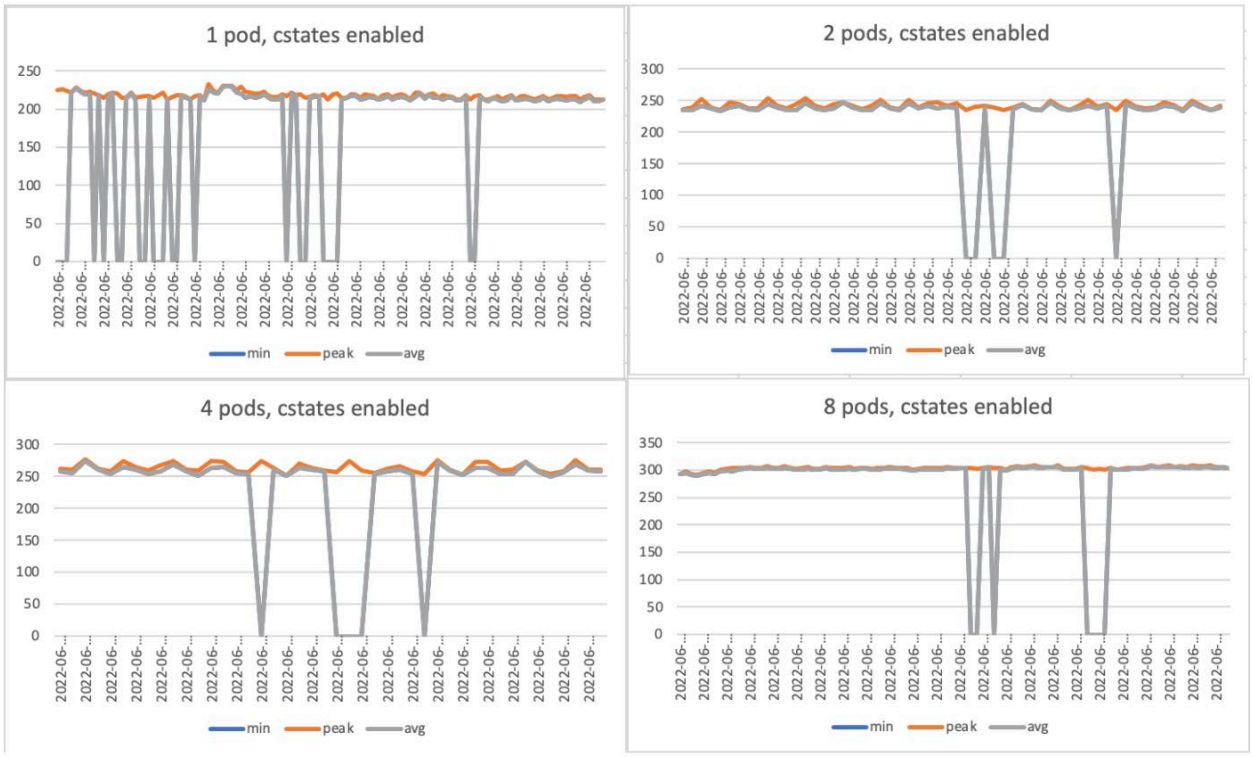


Figure 11 - Power Consumption of a Host with C-State Enabled[1]

With C-state enabled, as the number of CNF workloads deployed on a given host increases, the power consumption level rises. This is shown in Figure 11.

3.3. Future Considerations

The aggregate usage can be generalized as similar demand curve each day. An example of subscriber usage aggregated across all workloads within a PPOD is shown in Figure 12 and Figure 13. The Y-axis represent the workload usage percentage normalized by the access technology capacity service by the CNF. The X-axis is day of the month.

Figure 12 plots several diverse, individual workloads usage in a PPOD. Figure 13 is a stacked version, which provides a better view of the aggregate usage curve. The pattern is similar for all PPODs across all sites.

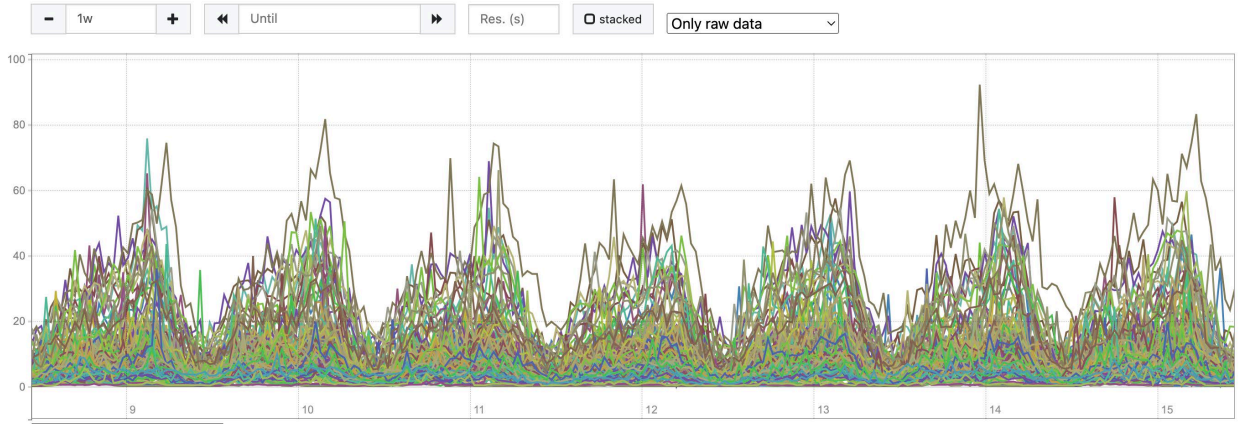


Figure 12 - Workload Utilization Percentage for a PPOD for One Week[1]

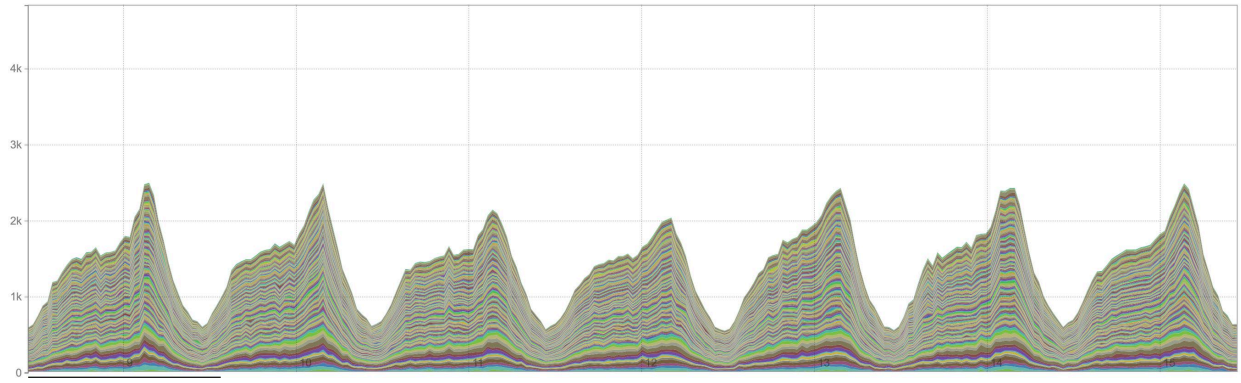


Figure 13 - Stacked Workload Utilization Percentage for a PPOD for One Week[1]

A system is elastic and can adapt to workload changes by provisioning and deprovisioning resources, in order to meet demand. On the compute cluster, significant CPU core resources are isolated and dedicated to the DPDK CNF workloads. The required CPU core resources for the CNF workloads are very much traffic or network IO bandwidth driven.

In the future, we will be exploring the following approaches to match the CPU resources to the traffic demand:

A. CPU P-State control by software application to match the short-term traffic demand
B. Bin packing of workloads across hosts to match the longer-term traffic demand
C. Combination of the above

Beyond the PPOD, we could also explore the overall capacity planning process in terms of spectrum activation. For example, an average 50-subscriber service group having low bandwidth demand requires far less spectrum and compute resources activated, as compared to that of an average 500-subscriber service group having much higher bandwidth demands. Can the capacity planning process be automatic, just-in-time, and elastic?

[1] *We collect, store, and use all data in accordance with our privacy disclosures to users and applicable laws.*

4. Conclusion

We started with energy consumption observability for the Comcast Private Cloud. We quantified the power savings of ~4.5% by reconfiguring the PSUs to hot standby; this will be operationalized for PPODs being put into production in late-2022/early-2023. As hot standby is platform specific, integrated software-based, user-defined configuration, we will be working with OEMs to explore how many existing platforms can be integrated through software upgrades. All hosts in PPODs are already provisioned with C-State power saving mode enabled.

The measured power consumption metrics for host with CPU C-State setting and various CNF workloads provide us the insight into the potential savings. The future strategy for energy efficiency is very much aligned with our cloud native architecture evolution; meaning it is just-in-time and elastic to workload changes by automatically provisioning and deprovisioning resources, such that the available resources match the demand.

Abbreviations

AC	Alternating current
BIOS	Basic input/output system
CCAP	Cable converged access platform
CIN	Converged Interconnect Network
CNF	Containerized network function
CPU	Central processing unit
CM	Cable modem
DAA	Distributed access architecture
DC	Direct current
DOCSIS	Data over cable interface specification
DPDK	Data plane development kit
DPoE	DOCSIS provisioning of EPON
EPON	Ethernet passive optical network
ELK	Elasticsearch, Logstash, and Kibana stack
iCMTS	Integrated cable modem termination system
HAGG	Headend aggregation switch
IP	Internet protocol
IPMI	Intelligent platform management interface
MAC	Media access control
MHAv2	Modular headend architecture version 2
OEM	Original equipment manufacturer
OLT	Optical line termination
ONU	Optical network unit
PHY	Physical
PPOD	Physical pod
PSU	Power supply unit
RF	Radio Frequency
RPD	Remote physical device
TSDB	Time Series Database
UR	Upstream router

vBNG	Virtual broadband gateway
vCM	Virtual cable modem
vCMTS	Virtual cable modem termination system

Bibliography & References

Dell PowerEdge Manuals

HPE ProLiant Manuals

Cable Labs R-PHY Specification

