

AIネットワークの可視性の提供

CVP UNO(Universal Network Observability)-

Shishio Tsuchiya shtsuchi@arista.com

ネットワーク管理者の苦悩



- ネットワークの変更で重要な機能が壊れる可能性はあるか？
- ネットワーク上でどのアプリケーションがどこで実行されているか？
- 仮想マシンが移動またはスケールアップしたことでの、ネットワークの信頼性やパフォーマンスに影響があったか？
- ネットワークのせいになってるが他はすべて正常に動作しているのか？
- 設定変更によってネットワークに問題が生じる可能性はあるか？
計画した変更は意図しない結果をもたらしたのか？

今のツールは「症状」の検出に優れているが
根本原因の特定は依然として手作業

Cloud Vision Universal Network Observability

従来の IT 環境



手動による診断

- データアクセス／データ洪水
- 過剰なコンソールウィンドウ
- 手動による処理と分析
- 遅いインシデント分析

最新のエンタープライズ



自己診断ネットワーク

- 集中および最適化されたステート (状態) の保管
- 単一画面
- クラウドベースの AI/ML 处理
- プロアクティブ リスク分析 / 迅速な インシデント分析 / 正確な 根本原因分析

CV UNO: CVaaS + Network Observability



自動デプロイメント

ゼロタッチ、ネットワーク全体の設定データモデル、
ビルト+カスタム Studiosワークフロー



リアルタイムテレメトリ

メトリクス、フロー、イベントのきめ細かな状態ス
トーリーミング、相関、傾向分析、予測アルゴリズ
ム、ネットワーク全体



変更管理

ネットワーク全体のアップグレード、ロー
ルバック、スナップショットのオーケスト
レーション



ネットワークの観測性

アプリケーションディスカバリー
物理/仮想ホストディスカバリー
アプリケーションダッシュボード
アプリケーション依存関係のマッピング
マルチドメインの影響分析
イベントの相関関係
問題の推論



コンプライアンス / リスク

インベントリ、逸脱、脆弱性、バグの継続的な評
価、報告、修正



セグメンテーションサービス

セキュリティポリシー管理 / 実施、ポリシー / アイデン
ティティ統合、ワイヤレス IPS



データセンター、キャンパス有線/WiFi、パブリッククラウド、WANインターネット

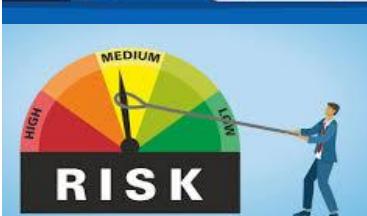


CloudVision



CloudVision

特にAIファブリックで観測可能性が重要な理由



AIワークロードは複雑

- 大規模なAIモデルは分散コンピューティングに依存しているため、問題検出が困難
- 性能のボトルネックは、ネットワーク、ハードウェア、ソフトウェアの非効率性の可能性

可視性の欠如＝リスクの増大

- 隠れた障害は、モデル精度の低下、遅延が増大し、非効率につながる
- リアルタイムの解析なしでAIパフォーマンス問題のデバッギングは、時間とコストがかかる

観測可能性がAIの最適化を解き放つ

- ネットワーク、システム、AIジョブのパフォーマンスをリアルタイムで深く可視化
- 問題の迅速な検出、根本原因の分析、プロアクティブな解決が可能
- GPU/TPUの利用率を最大化し、コストの削減と効率の向上

ビジネスインパクト

- モデル学習と推論の高速化＝市場投入までの時間の短縮
- リソースの最適利用による運用コストの削減
- AIの信頼性、拡張性、および長期的な成功を保証

AIファブリックは信頼性が重要

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

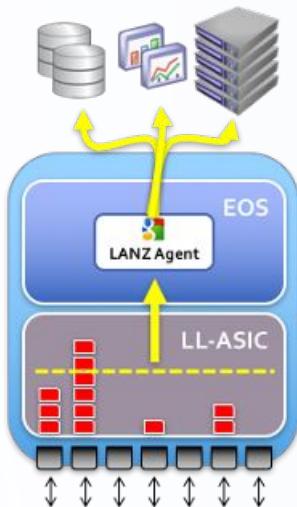
- MetaがLlama 3 405Bモデルのトレーニングを54日間にわたって実施
- この期間中、クラスターは合計466回のジョブ中断を経験
- 47回は計画的なメンテナンスだったが、残りの419回は予期せぬ障害によるものだった

https://www.tomshardware.com/tech-industry/artificial-intelligence/faulty-nvidia-h100-graphics-and-hbm3-memory-caused-half-of-the-failures-during-llama-3-training-one-failure-every-three-hours-for-metas-16384-gpu-training-cluster?utm_source=twitter.com&utm_campaign=socialflow&utm_medium=social

Latency Analyzer (LANZ)とCVPによる可視化

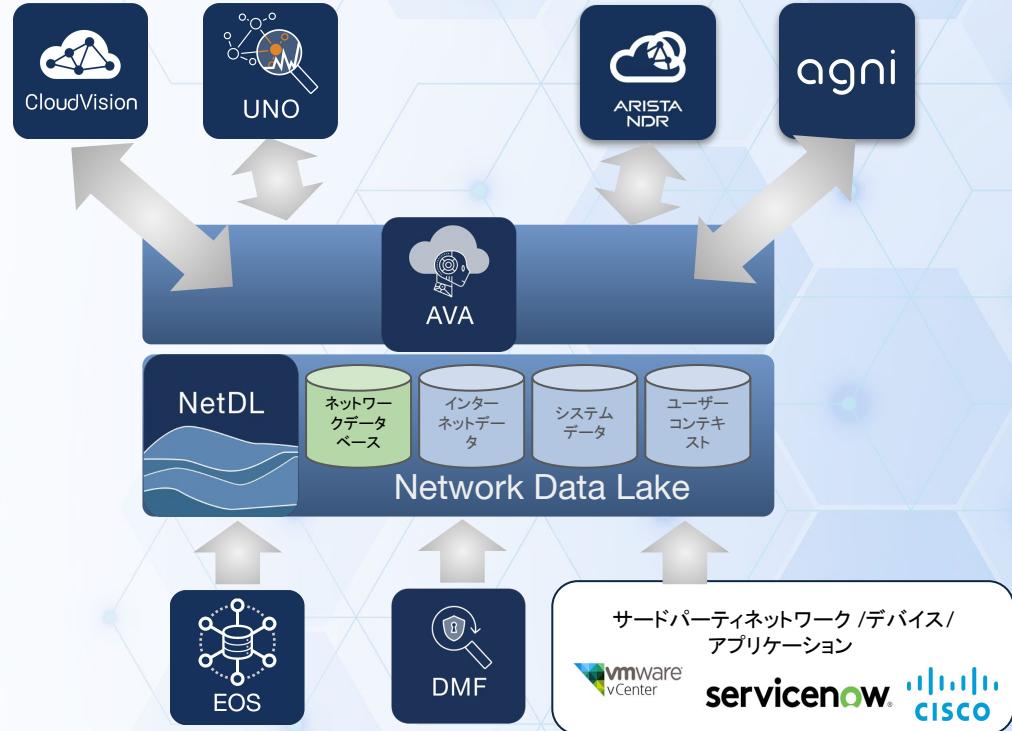


- EOS LANZ (Latency Analyzer):
 - マイクロ秒でキュー混雑、ドロップを記録
 - CVP:
 - 時系列で LANZ 情報を表示



データドリブンアーキテクチャにおける Network Observability

データレイクとは構造化されたデータ/非構造体データ/半構造体データを集めた中央のレポジトリ

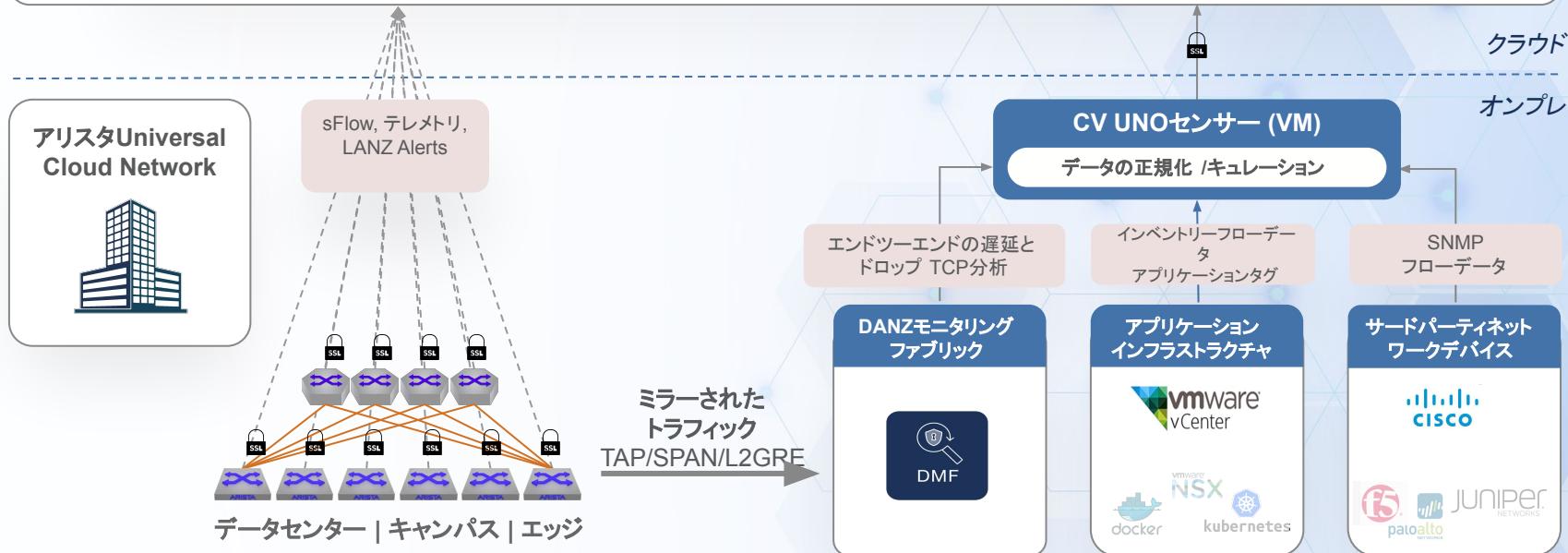


CV UNO アーキテクチャモデル



CVaaSとCV UNO

アリスタのホスティング /
マネージドインフラストラクチャ



CV UNO 根本原因の分析：ネットワークの推論

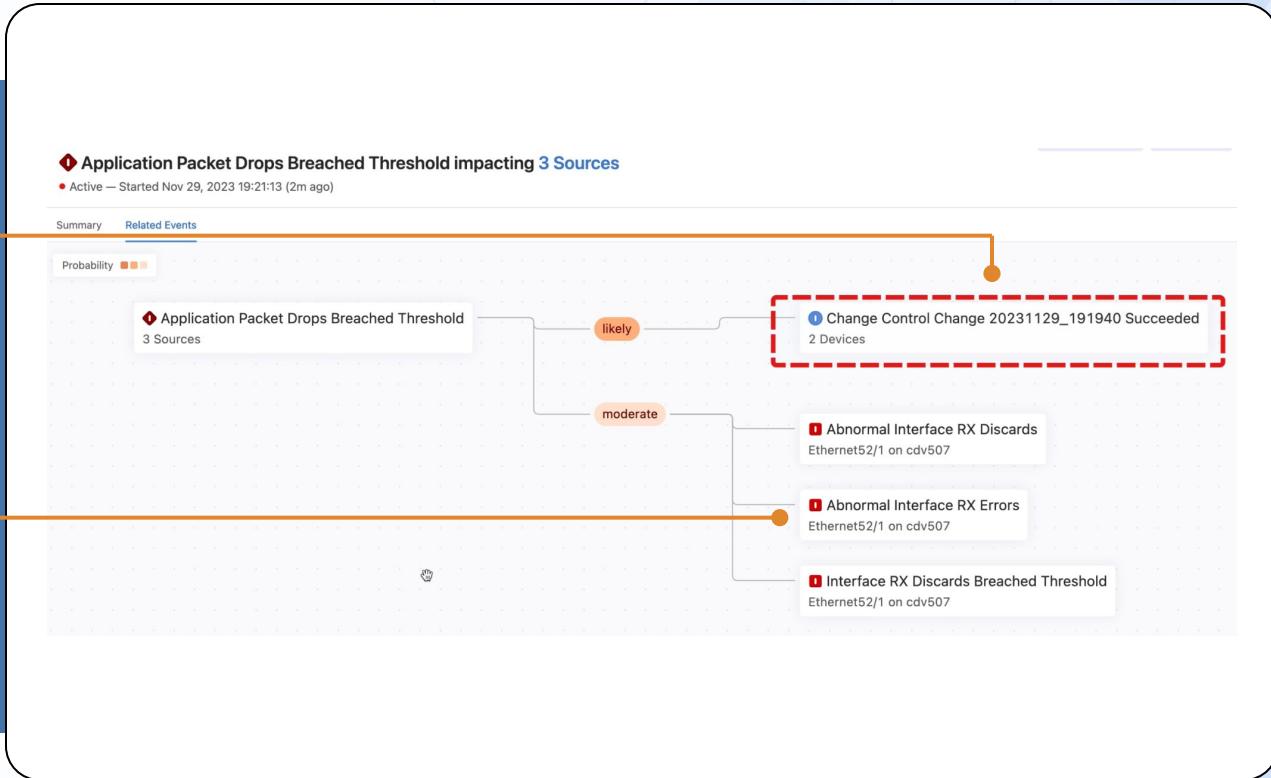


根本原因の分析

問題の引き金となる
ネットワーク変更の可
能性を特定

根本原因の分析

ネットワークパフォーマ
ンスイベントが問題を
引き起こす可能性は
低い



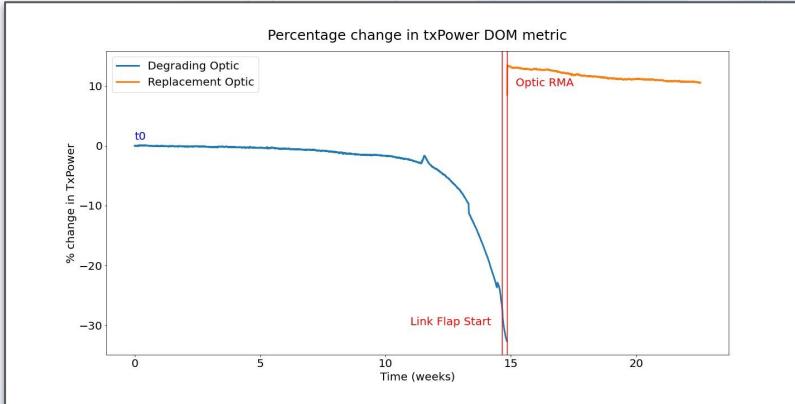
コンポーネント故障の事前予知



BETA

オプティクス障害の例

- リンクの送受信両方のDOMデータをNetDLが蓄積し、AVAが分析
- 電流、温度、フレームエラー、送受信光強度に基づき、AVAがオプティクスの故障をモデル化
- AVAは、実際に障害が発生する前に、ユーザやアリスタにアラートをあげ、プロアクティブなオプティクスのRMAを実施



- エンジニアリングチームは2つの大規模顧客と緊密に連携
- CVPIにより、22Kのオプティクスを1年間にわたって監視
- 16件の障害が予測され、その後、その障害が確認された(0誤検知)
- 現在、すべてのCVaaS顧客に導入中
- TACによるプロアクティブなRMAの開始

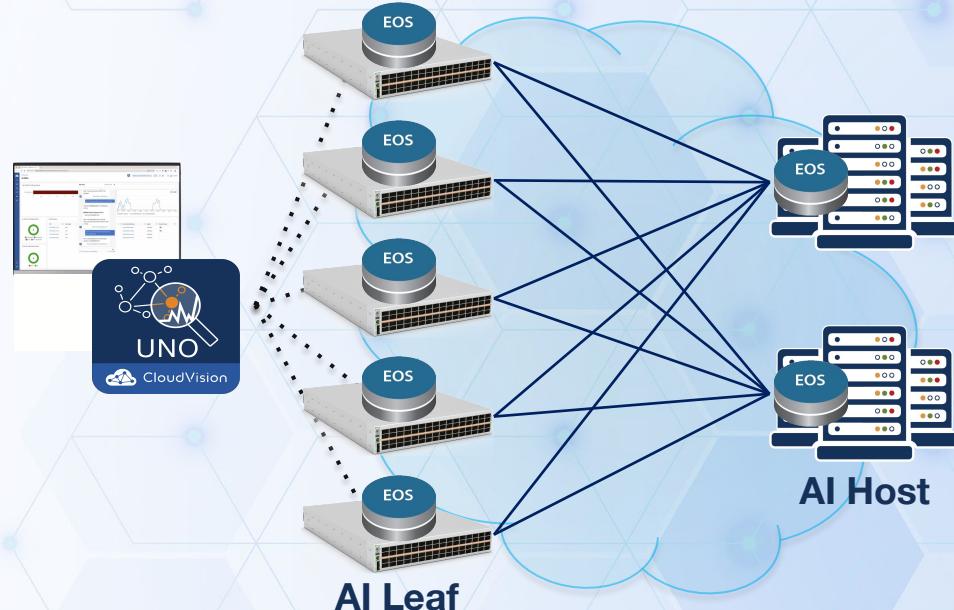
AI向けのCVP UNOの機能

1. Jobとネットワークのパフォーマンスデータの統合
2. 高精度ストリーミングデータとサードパーティデータの統合
3. Jobパスの可視化
 - a. サーバー/NIC/ネットワークに横断するパスを可視化
4. 相関分析と推論
 - a. AIを活用したイベント相関分析や異常/障害検出



CV UNOの利点

- **全体的なAIジョブ監視** – ジョブの健全性、輻輳、リソース利用状況をマイクロ秒単位でリアルタイムに把握
- **ディープダイブ分析** – ネットワークデバイス、サーバーNIC、RDMAエラーを分析することでボトルネックを特定
- **エンド・ツー・エンド・フローの可視化** – AIによる直感的なジョブ・フロー追跡により、問題を迅速に検出
- **プロアクティブな問題解決** – 早期の異常検知により、AIワークロードの実行を中断を防ぐ



CV UNO Dashboards: AI Job View

AI Jobs

Active Jobs: 24 Total Servers: 680 Total NICs: 5440 Total Leaf Switches: 170 Total Flows: 88064

Last 30 Days

Jobs



Drops



PFC (RX)



PFC (TX)



Queue Utilization over Threshold Events



Total ECN



Total Marked ECN



24 Jobs

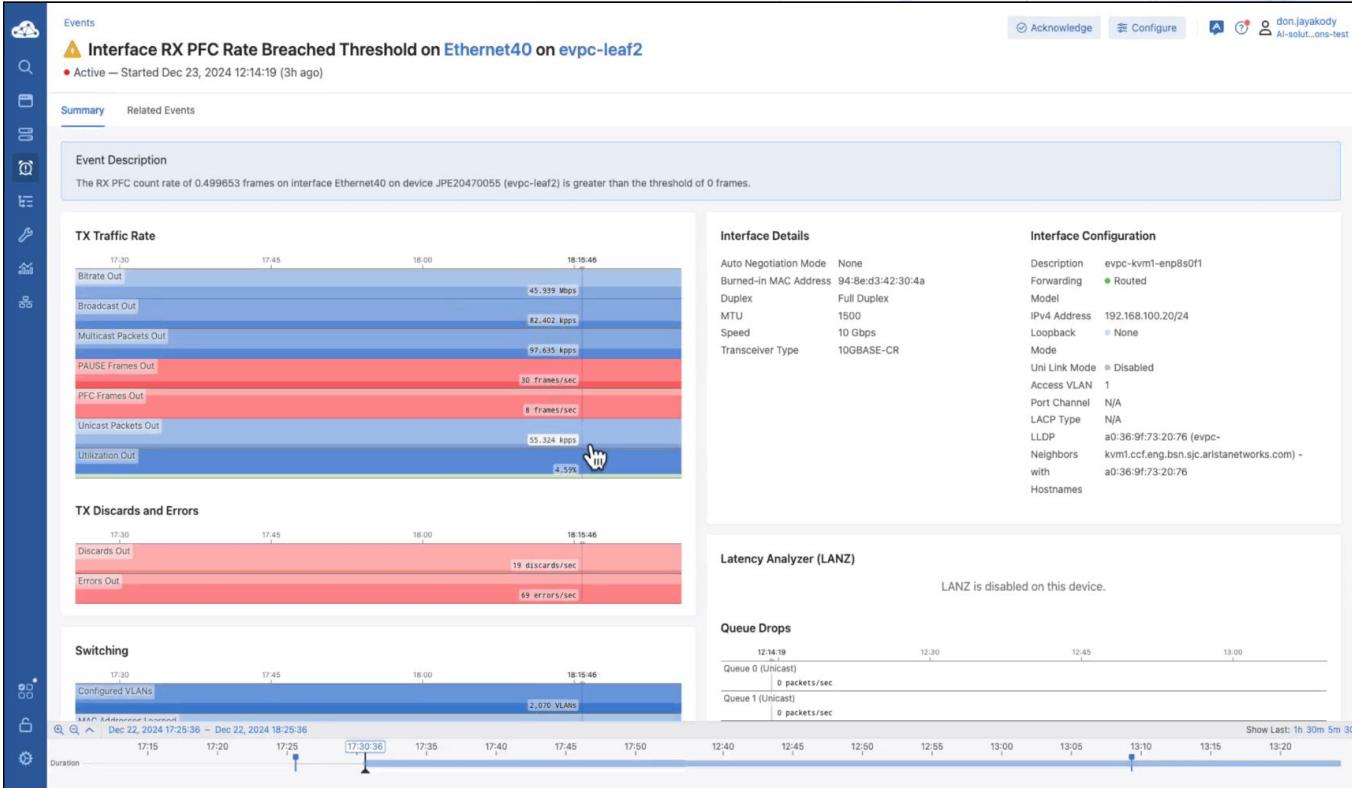
Job ID	Start Time	End Time	Duration	Lea...	Servers	NICs	Flows	Avg Ntw Util	Status	Health	Events
12345	Oct 28, 2024 09:21:45	Oct 29, 2024 07:42:28	22hr 21min	16	64	512	8192	88%	● Failed	● Critical	● 2, ● 4, ● 8
12346	Oct 27, 2024 10:47:17	Oct 27, 2024 16:47:20	6 hr 1min	14	56	448	7168	75%	● Failed	● Critical	● 1, ● 4, ● 1
12347	Oct 27, 2024 08:43:05	—	—	13	52	416	6656	90%	● Running	● Healthy	● 4, ● 1
12348	Oct 26, 2024 09:21:45	—	—	11	44	352	5632	84%	● Running	● Healthy	● 2, ● 1

AIのJob IDの実行時間やステータスが表示

Failになった12345というjobに関してドリルダウン

サーバー/Leaf Switch/Flowとドリルダウン解析が可能

CV UNO Dashboards: Events for AI



leaf2のRXでPFCを受けている。他のインターフェースを見るとECNのスレッシュホールドに到達設定の見直しを実施

CloudVision Dashboards: Network Health

The access permissions assigned to your account do not allow you to configure this section. Contact your CloudVision administrator in relation to access permissions.

Jan 10, 2025 14:33:20 (1 hour) Edit ... Cloud Staging Cluster praful.bhaidasna AI-solutions-test

Network Health

Overview of the health metrics. #built-in

Devices by Category

Category	Count
Hardware	1
Environment	10
L2	2
Routing	2
Security	2
Compliance	1
Network Fabric	10

Legend: Critical (red), Error (orange), Warning (yellow)

Top Network Event Types

Event Types	Category
Unexpected link state (89 Times 12 Devices)	L2
Device Power (43 Times 27 Devices)	Environment
CVE Threat Event (34 Times 22 Devices)	Security
Streaming Activity (30 Times 30 Devices)	Network Fabric, CVP
Hardware Failure (24 Times 12 Devices)	Network Fabric

Showing 10 items

Streaming Information

37 Devices

Active: 35 Inactive: 2

Devices | 37

Hide Devices with No Events

Group by: Pods Sort: Health

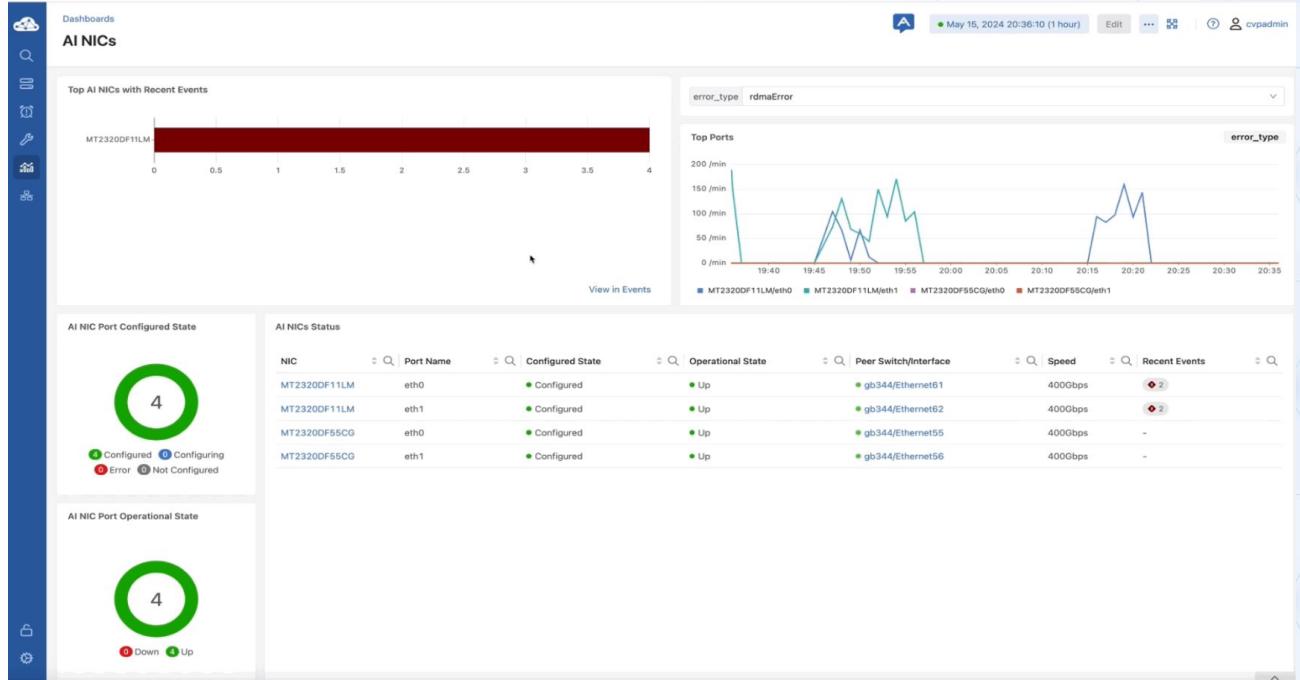
Legend: Critical (red), Error (orange), Warning (yellow), Healthy (green)

Device	Health
ld568-spine18	Healthy
ld568-spine18	Healthy
lmf359-spine2	Healthy
mty333-spine17	Healthy
ka151-Spine-2-nfc468-Spin...	Warning
Unclassified	Healthy
Untagged	Healthy

ネットワーク全体の各デバイスの時間ごとのCritical/Warning/info情報のヒートマップを表示

スイッチ/サーバとそれぞれの詳細ドリルダウンが表示可能

CV UNO Dashboards: NIC Health



NICのAI特有のイベントのダッシュボードを表示

RDMAエラー/PCIeエラー/インターフェースエラーやPFC/ECNカウンター

QoSの設定などを一括表示

CV UNO Dashboards: Simplified Incident Analysis



NIC Details

device: MT2320DF11LM

NIC Details

Model	Version	Machine	Host Name	Chassis ID	Kernel Name	Serial Number	Kernel Release	Kernel Version
BlueField-3 SmartNIC Main Card	1.0.0	aarch64	NicDocaHost1	58a2.e100.0100	Linux	MT2320DF11LM	5.15.0-1022-bluefield	#24-Ubuntu SMP Mi

Port & Peer Details

Port	Configured State	Operational State	Peer Switch/Interface	MAC	MTU	Speed	Duplex
eth0	Configured	Up	gb344/Ethernet61	58a2.e100.01.01	9000	4000bps	Full Duplex
eth1	Configured	Up	gb344/Ethernet62	58a2.e100.01.02	9000	4000bps	Full Duplex

RDMA Error Rate

eth0: RDMA Error Rate (0/min to 150/min) over time (19:40 to 20:30). A sharp peak is visible around 19:50.

eth1: RDMA Error Rate (0/min to 200/min) over time (19:40 to 20:30). A sharp peak is visible around 19:50.

Ask AVA

Here is the dashboard to all information relevant to MT2320DF11LM
[View MT2320DF11LM Dashboard](#)

What could have caused the RDMA event on eth1 on MT2320DF11LM?

There are 2 Related Events associated with RDMA Packet Sequence Error on AI Port1 on AI NIC1.

Related Event 1: **Interface TX Discards Breached Threshold**
Ethernet61 on leaf-c5-0

Related Event 2: **Queue Size Breached Threshold**
Ethernet61 on leaf-c5-0

Here is the page that shows related events.
[Related Events](#)

Type to ask me something

Ask AVAにエラーの詳細原因を尋ねることで、関連イベントを調査各種イベントを相関付けし、特定してくれる。トラブルシューティングが簡単に

ECMPの可視化



Traffic Overview



Devices

- Inventory
- Endpoint Overview
- Network Entities
- Device Registration
- Compliance Overview
- Endpoint Overview
- Connectivity Monitor
- Traffic Overview
- Traffic Flow
- Comparison

Multi-Cloud Dashboard

Network Segmentation

Search

Live Time

?

arista

roseebyrne

Traffic Overview

Network Hierarchy: All | Device Type: All

Total Uplink Utilization

TX: 63% (1.04 Tbps) Total Bandwidth: 1.58 Tbps

RX: 98% (1.48 Tbps) Total Bandwidth: 1.58 Tbps

Total Downlink Utilization

TX: 94% (1.74 Tbps) Total Bandwidth: 1.98 Tbps

RX: 78% (1.04 Tbps) Total Bandwidth: 1.98 Tbps

Aggregate Traffic

TX and RX | TX | RX

254 Devices

Sort by: Uplink TX Balance (High to Low) | Edit Columns

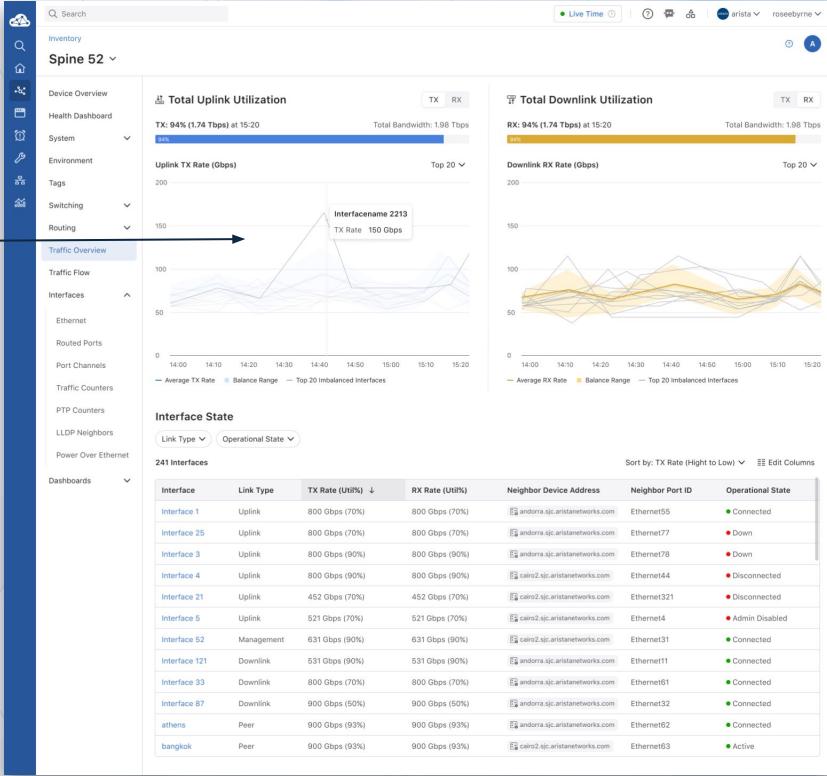
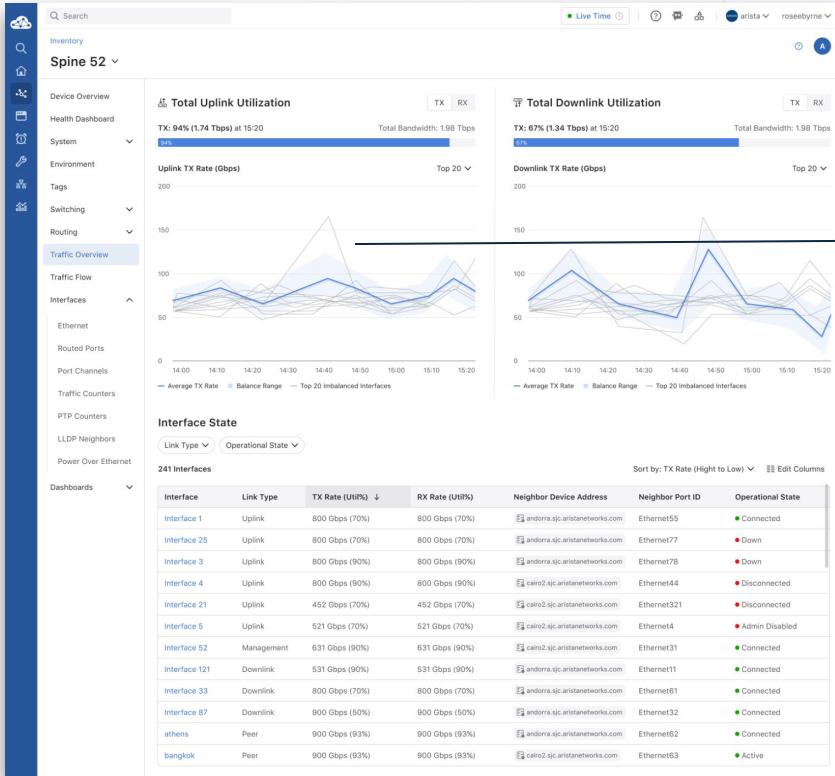
Device	Direction	Uplink Balance	Uplink Rate (Util%)	Number of Uplinks	Downlink Balance	Downlink Rate (Util%)	Number of Downlinks
Spine 52	TX	80.53 Gbps	800.22 Gbps (70.4%)	28	80 Gbps	800 Gbps (70%)	28
	RX	53.33 Gbps	1.03 Kbps (70.3%)	28	53 Gbps	800 Gbps (70%)	28
beagle	TX	40 Gbps	800 Gbps (90%)	20	40 Gbps	800 Gbps (90%)	20
	RX	30 Gbps	800 Gbps (90%)	20	30 Gbps	800 Gbps (90%)	20
berlin	TX	0 Gbps	452 Gbps (70%)	24	0 Gbps	452 Gbps (70%)	24
	RX	0 Gbps	521 Gbps (70%)	24	0 Gbps	521 Gbps (70%)	24
belem	TX	0 Gbps	631 Gbps (90%)	100	0 Gbps	631 Gbps (90%)	100
	RX	0 Gbps	531 Gbps (90%)	100	0 Gbps	531 Gbps (90%)	100
bii234	TX	0 Gbps	800 Gbps (70%)	44	0 Gbps	800 Gbps (70%)	44
	RX	0 Gbps	900 Gbps (50%)	44	0 Gbps	900 Gbps (50%)	44
bii234	TX	0 Gbps	900 Gbps (93%)	24	0 Gbps	900 Gbps (93%)	24
	RX	0 Gbps	900 Gbps (93%)	24	0 Gbps	900 Gbps (93%)	24

21 | Copyright © Arista 2025. All rights reserved.

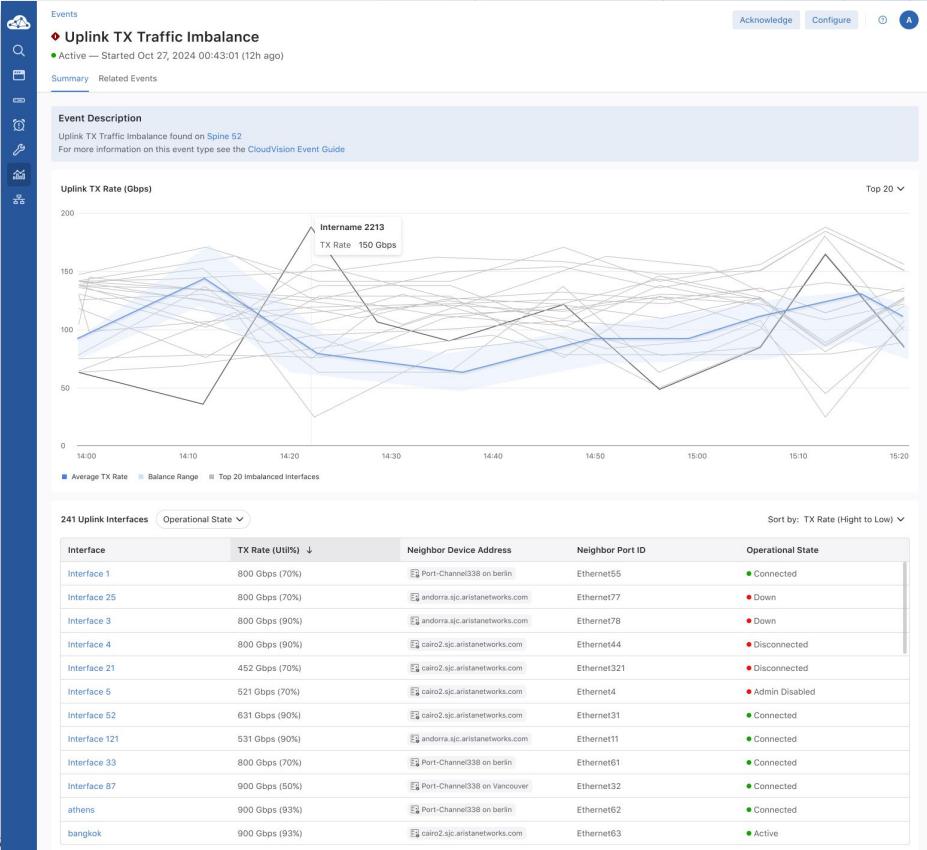
Public.

ARISTA 21

ECMP Per Device Traffic Overview

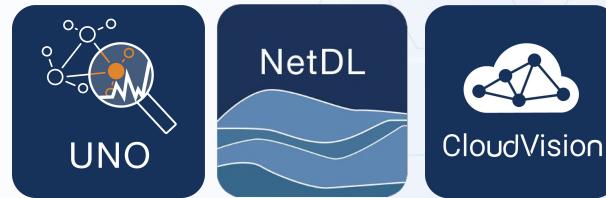


ECMP Imbalance Events



まとめ

- ネットワーク全体でロスレス/バーストで分散処理を行うAIファブリックにおいて必要と思われる可視性を提示
 - QoS設定およびカウンター(DSCP/PFC/ECN)
 - NIC/サーバーのカウンター
 - ネットワーク全体でのフローの相関付
 - JCTとJob IDとそれらの関連付け
 - ECMPの偏りの確認





Thank You

www.arista.com



FAQ

Q.LANZデーターの更新の頻度は？

A.LANZ自身はマイクロ秒単位で観測、ストリーミングテレメトリーとしてインターフェースデータなどは2秒間隔/LANZの閾値を超えたときにはsub secでのアップデートを行う。

Q.テレメトリーデーターの保管期間は？

A.生のデータは2秒毎に送られ集約データは10秒、1分、15分の各間隔で計算される。保管はCVaaSでは90日 / オンプレ版では7日間

Q.AI Jobデータやサーバーデータはどこと連携するか？

A.Slurm APIなどを通じてセンサーが情報を収集、サーバーに関してはPrometheusを使用

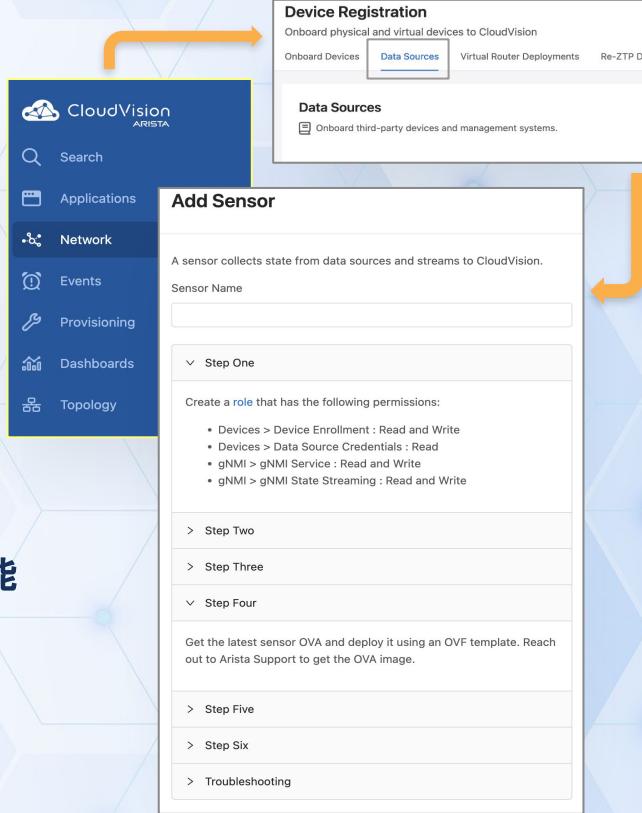
FAQ

Q.コンプライアンスダッシュボードはどの頻度でアップデートされるか？

A.AristaからAlertBaseというバグ情報/脆弱性を公開している。この情報を自動的に同期し、自分が管理するデバイスに対してチェックを行う。

CV UNO センサーのハイライト

- ✓ 様々なデータソースからデータを収集
 - 例: vCenter, ベアメタルサーバー, サードパーティーネットワークデバイス, DMF
- ✓ センサーは収集したデータを安全に CVaaSにストリーミング
- ✓ VMware環境の仮想ホストとして展開 (OVAアプライアンス)
- ✓ センサーは顧客施設内に設置
- ✓ ほとんどの場合、単一の CV UNOセンサーで DC全体に対応可能
- ✓ TLS 1.3 over HTTP2



CV UNO データソース

ネットワーク中心のデータ

EOS
(テレメトリーとフローデータ)



サードパーティー
(SNMP)



vCenter
(vDSwitchからのNetflow)



サードパーティー
(フローデータ)



アプリケーション中心のデータ

vCenter
vCenterインベントリデータ
(ESXi, VMs, vDS)



ペアメタル
帯域幅(接続されたスイッチインターフェースから得られる)



アプリケーション定義
(CMDBを活用)



パケットの分析

DMFサービスノード
(TCP分析 - エンドツーエンド遅延)

