ARISTA

AIネットワークの可視性の提供

CVP UNO(Universal Network Observability)-

Shishio Tsuchiya shtsuchi@arista.com

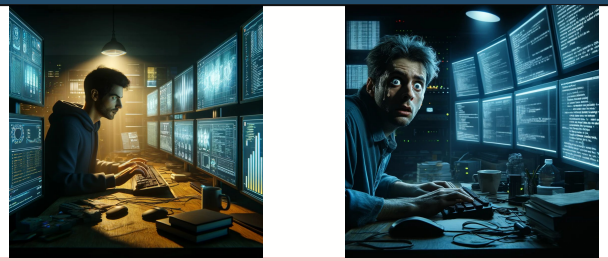
ネットワーク管理者の苦悩

- 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、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

UNO　NetDL　CloudVision

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処理**
- **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、プリビルド+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

CloudVision

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インターコネクト

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

ARISTA

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%

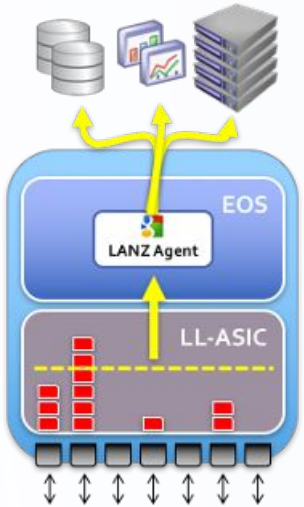
Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

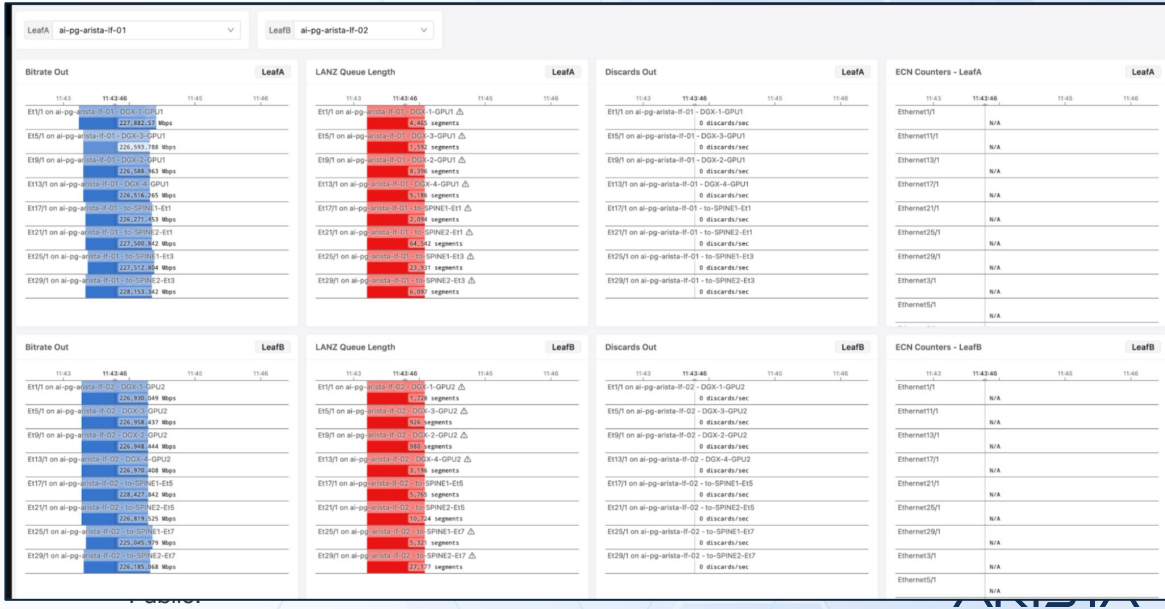
https://www.tomshardware.com/tech-industry/artificial-intelligence/fau
lty-nvidia-h100-gpus-and-hbm3-memory-caused-half-of-the-failures-d
uring-llama-3-training-one-failure-every-three-hours-for-metas-16384-
gpu-training-cluster?utm_source=twitter.com&utm_campaign=socialfl
ow&utm_medium=social

- 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

Latency Analyzer (LANZ)とCVPによる可視化

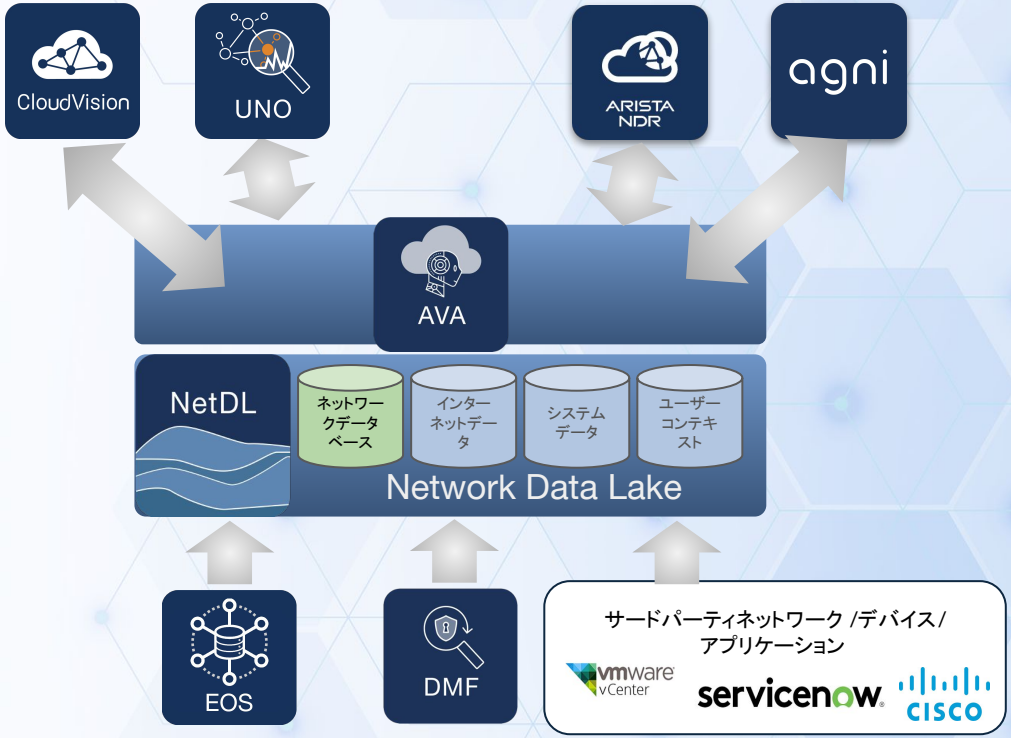
- EOS LANZ（Latency Analyzer):
  - マイクロ秒でキュー混雑、ドロップを記録
- CVP:
  - 時系列でLANZ情報を表示





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

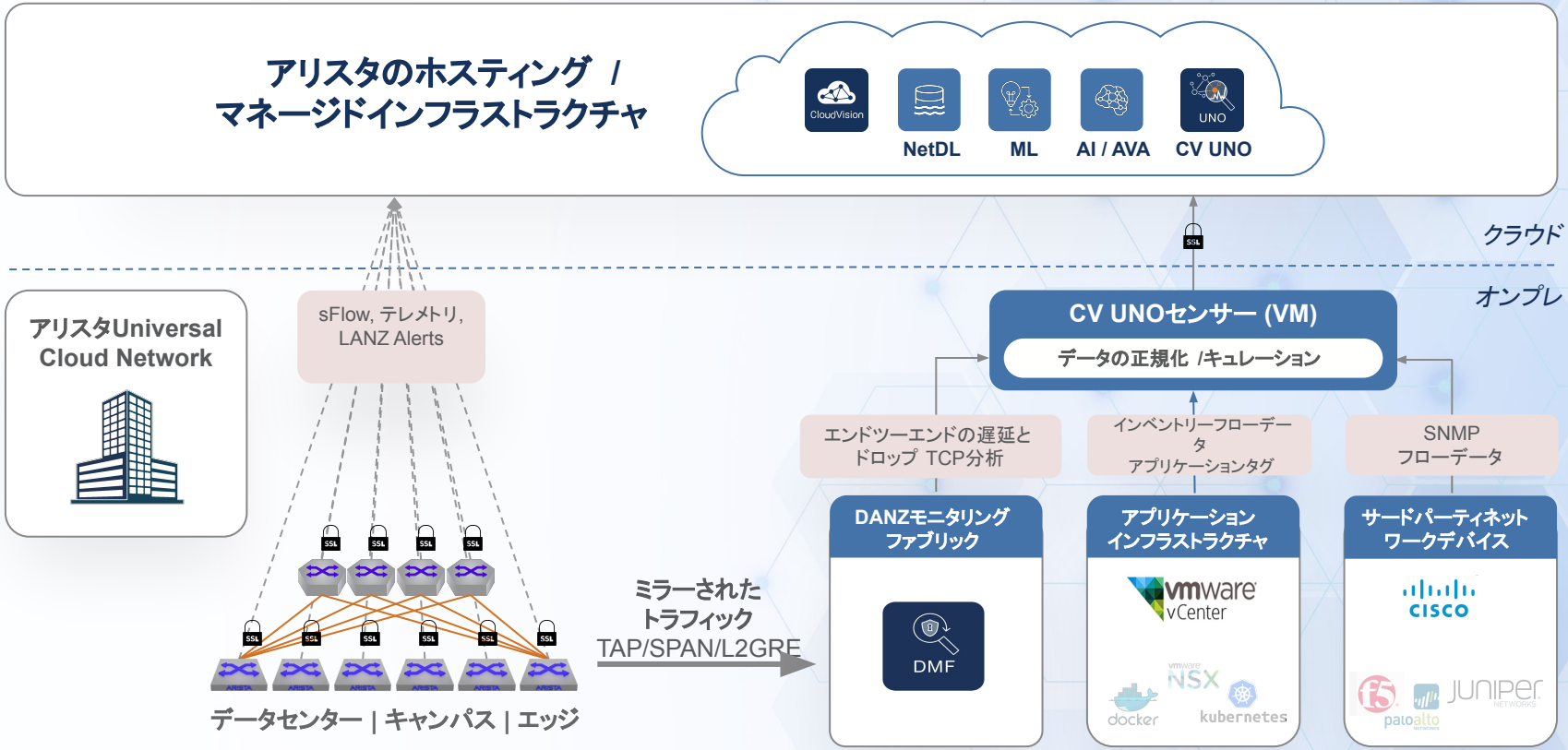
データレイクとは構造化されたデータ/非構造体データ/半構造体データを集めた中央のレポジトリ



ARISTA

CV UNO アーキテクチャモデル

CVaaSとCV UNO



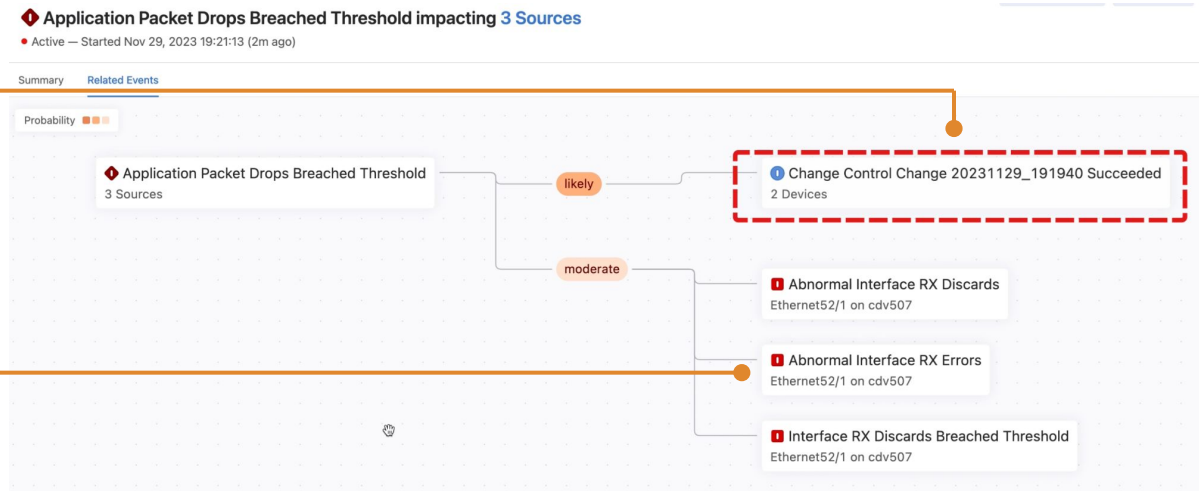
CV UNO 根本原因の分析：ネットワークの推論

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



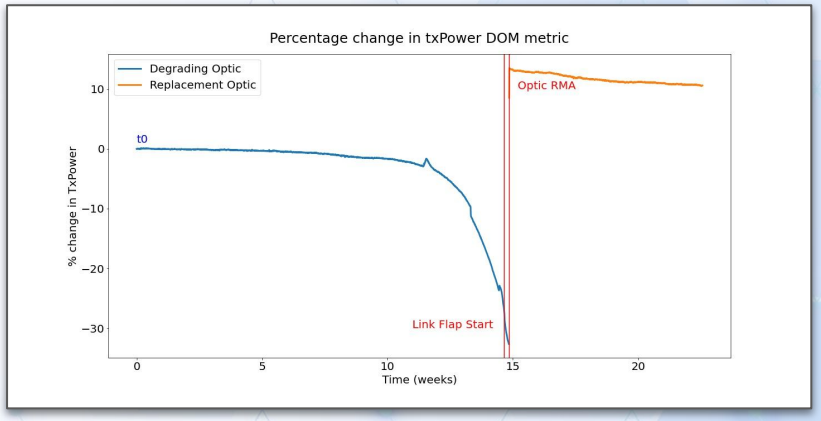
コンポーネント故障の事前予知

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

AVA- Autonomous Virtual Assist
Proactive NetOps
Predictive component failures
Device capacity planning and resource mgmt



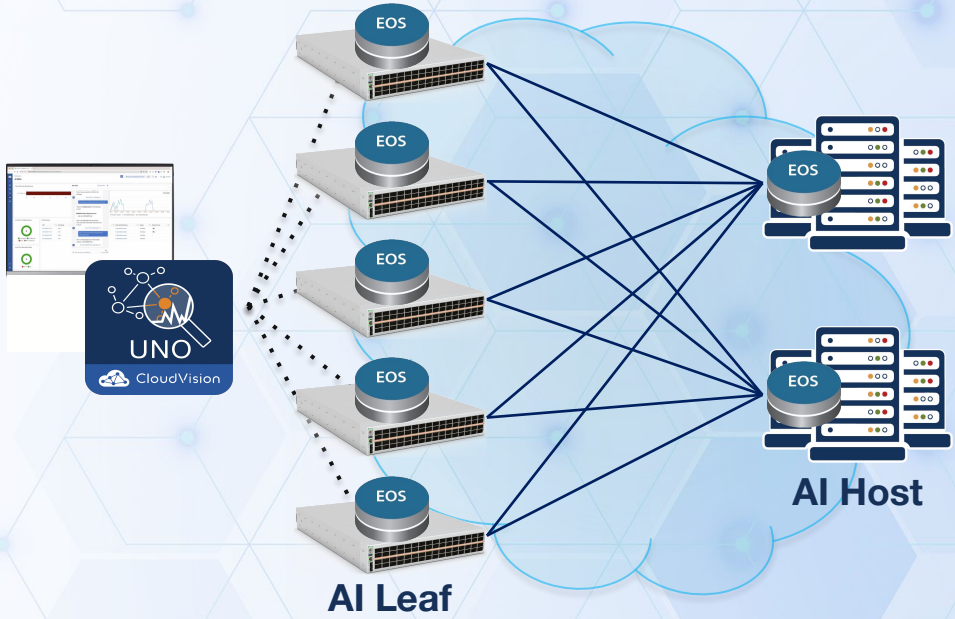
- エンジニアリング チームは 2 つの大規模顧客と緊密に連携
- CVPにより、22Kのオプティクスを1年間にわたって監視
- 16 件の障害が予測され、その後、その障害が確認された(0 誤検知)
- 現在、すべての CVaaS 顧客に導入中
- TAC によるプロアクティブなRMA の開始

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

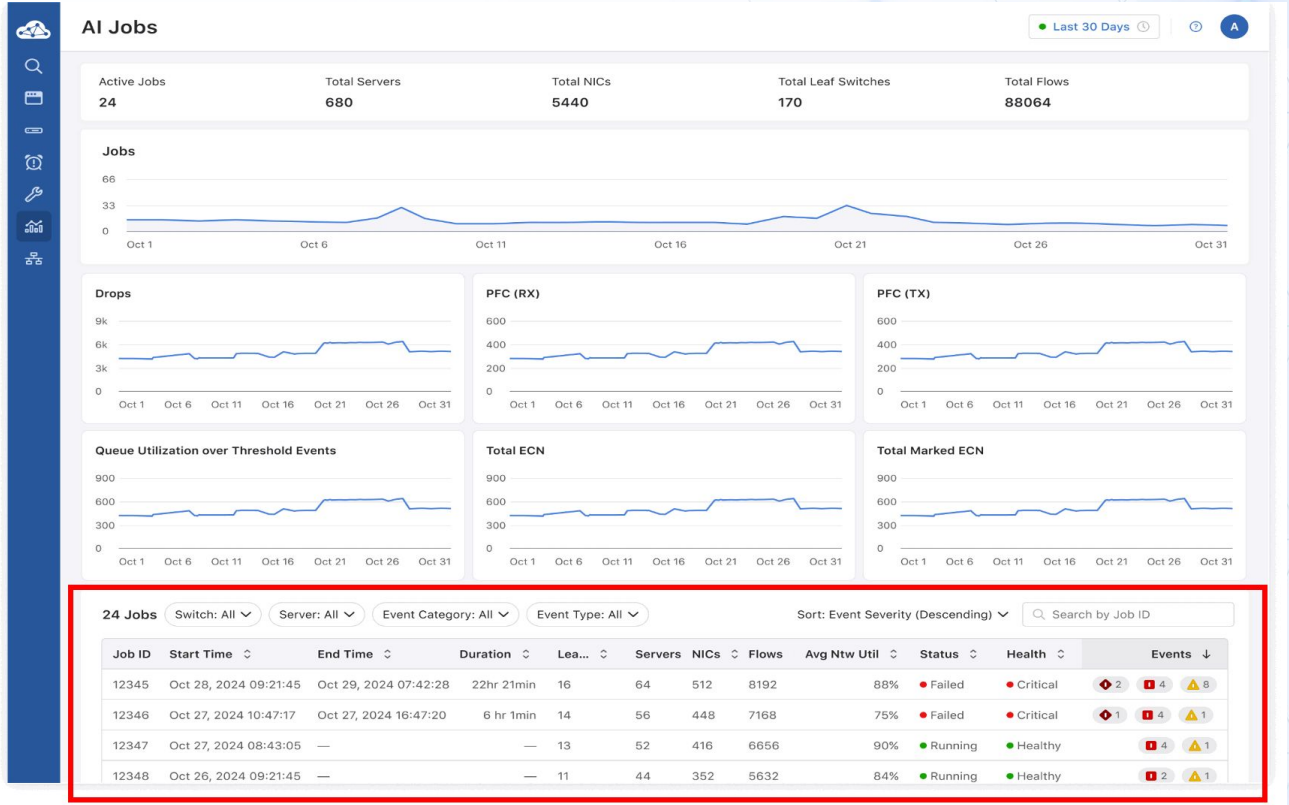
CV UNOの利点

- **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



ARISTA

CV UNO Dashboards: AI Job View

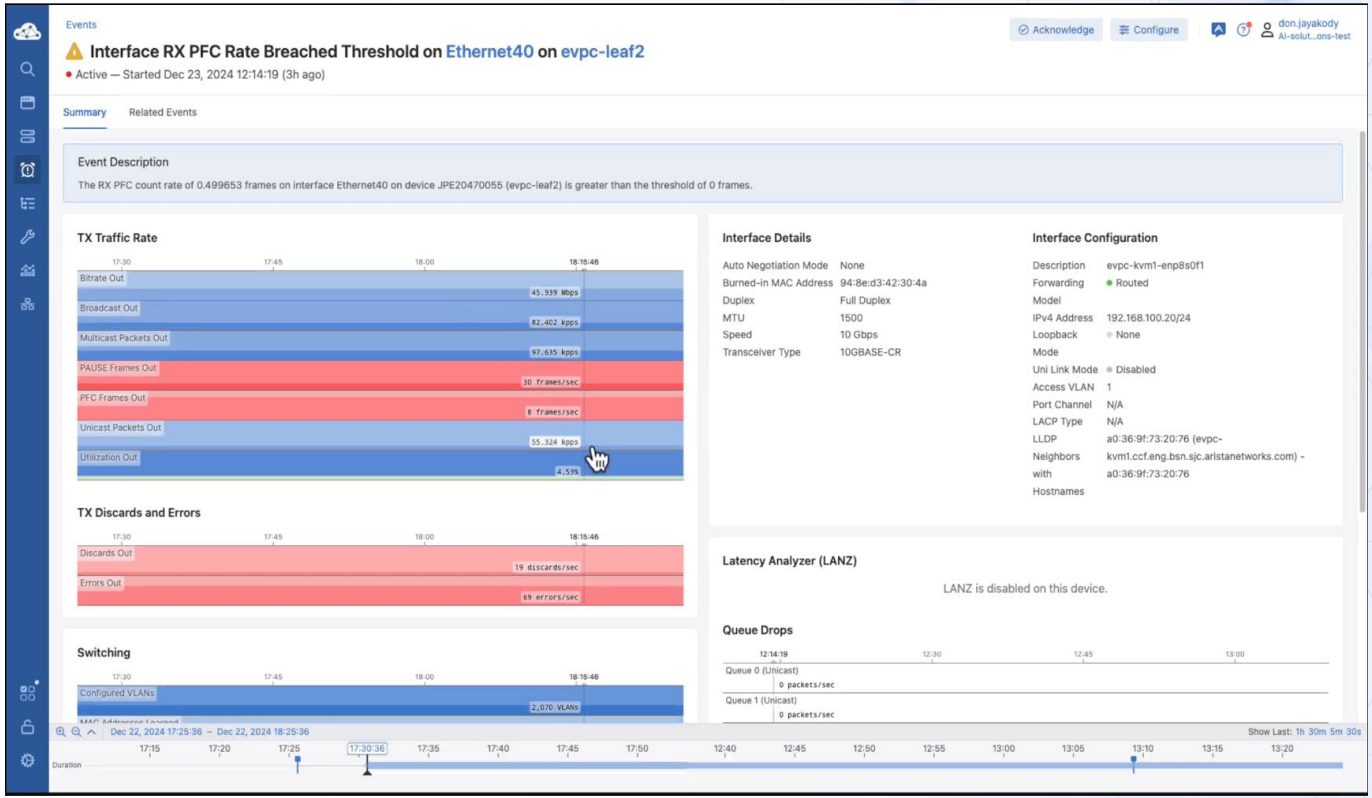


AIのJob IDの実行時間やステータスが表示

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

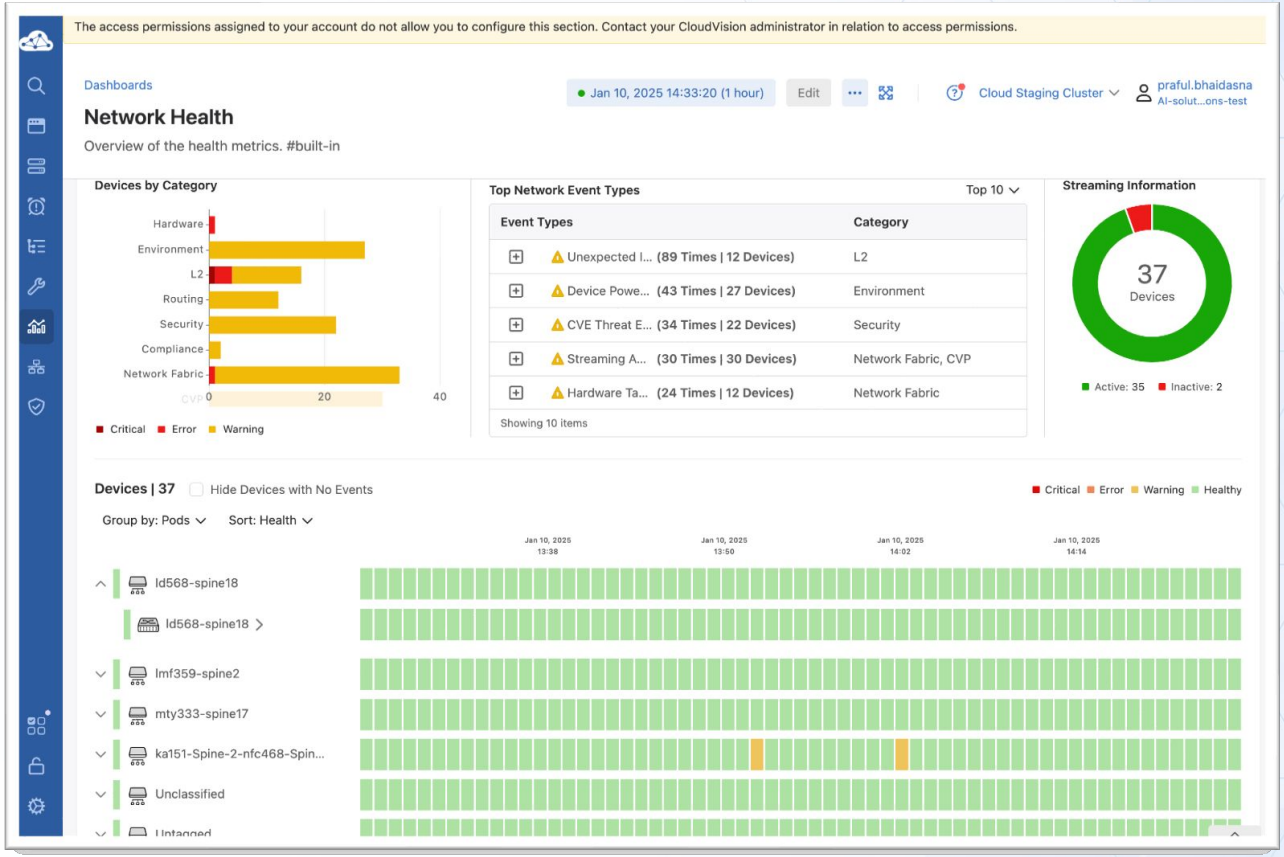
CV UNO Dashboards: Events for AI



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

Public.

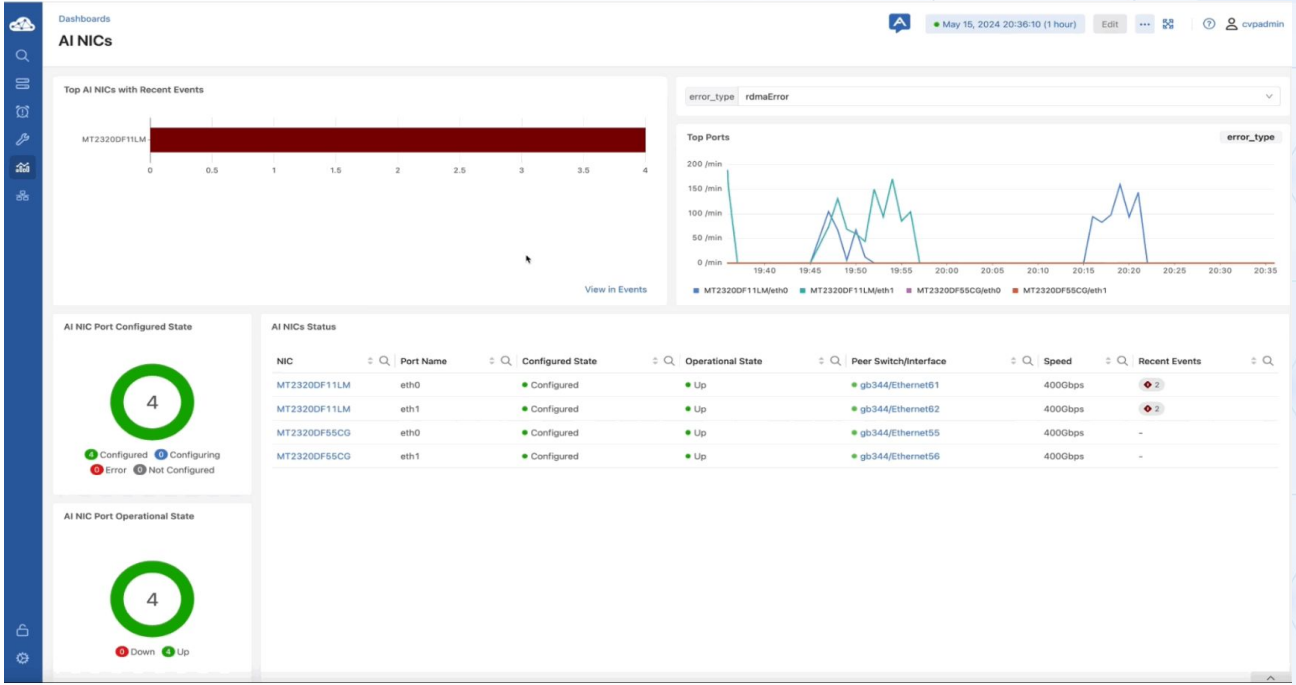
CloudVision Dashboards: Network Health



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

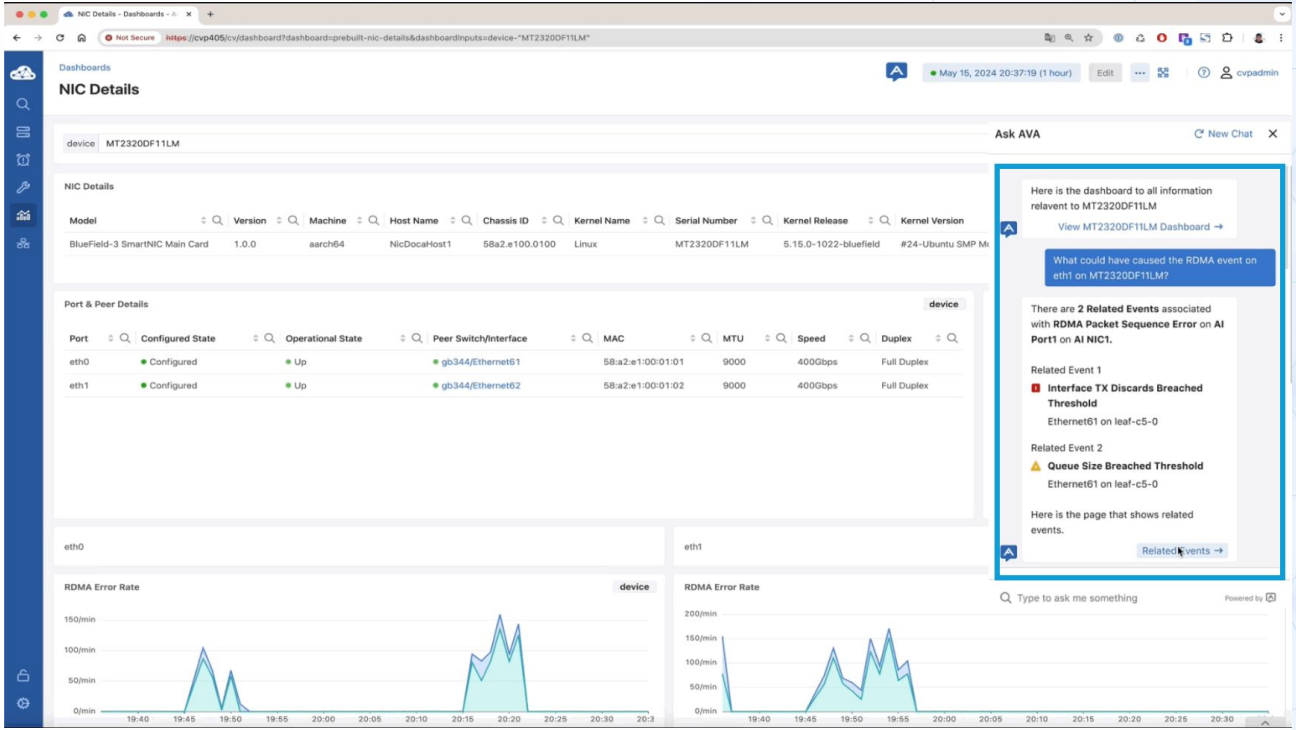


NICのAI特有のイベントのダッシュボードを表示

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



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

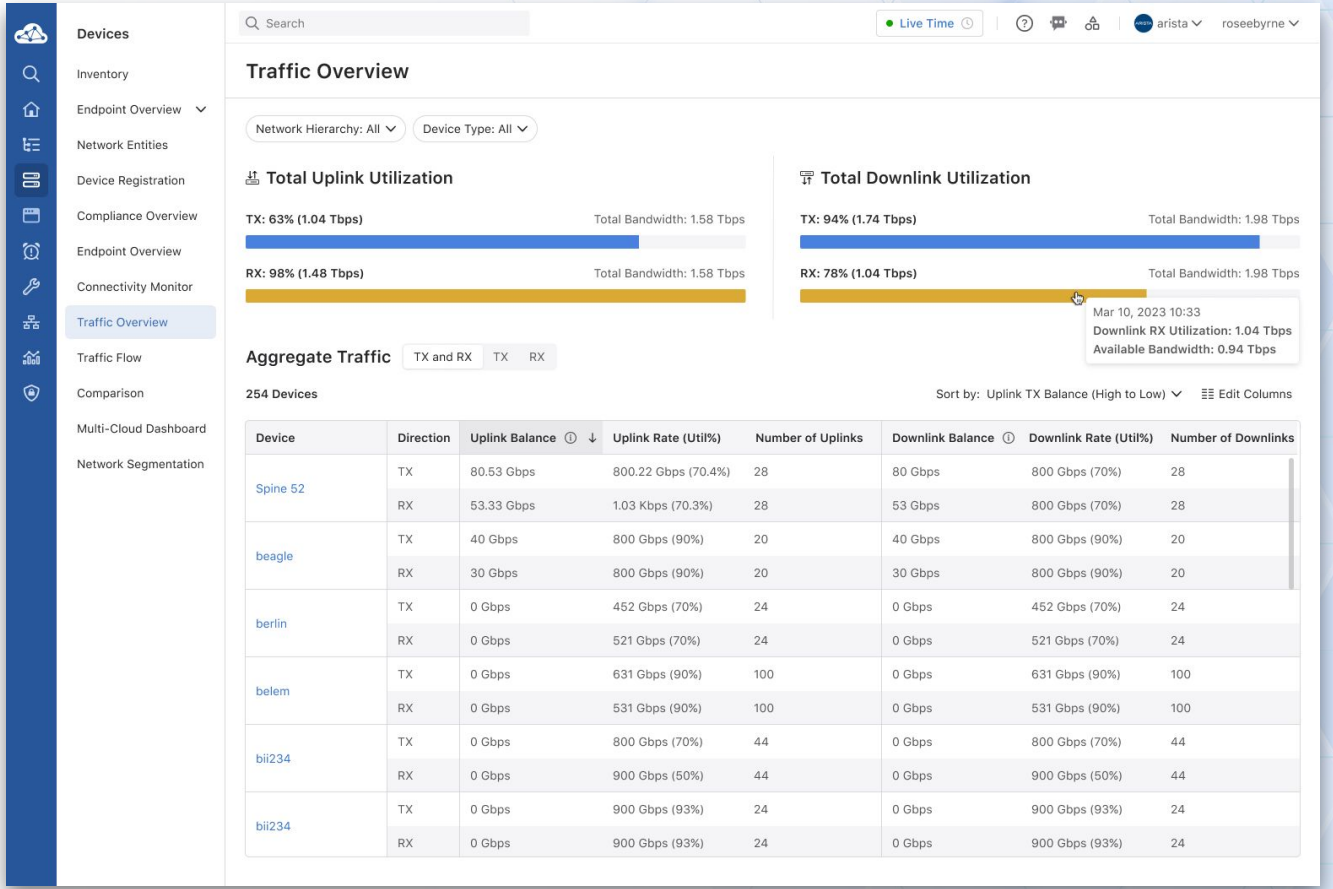
各種イベントを相関付けし、特定してくれる。

トラブルシューティングが簡単に

ECMPの可視化

Traffic Overview

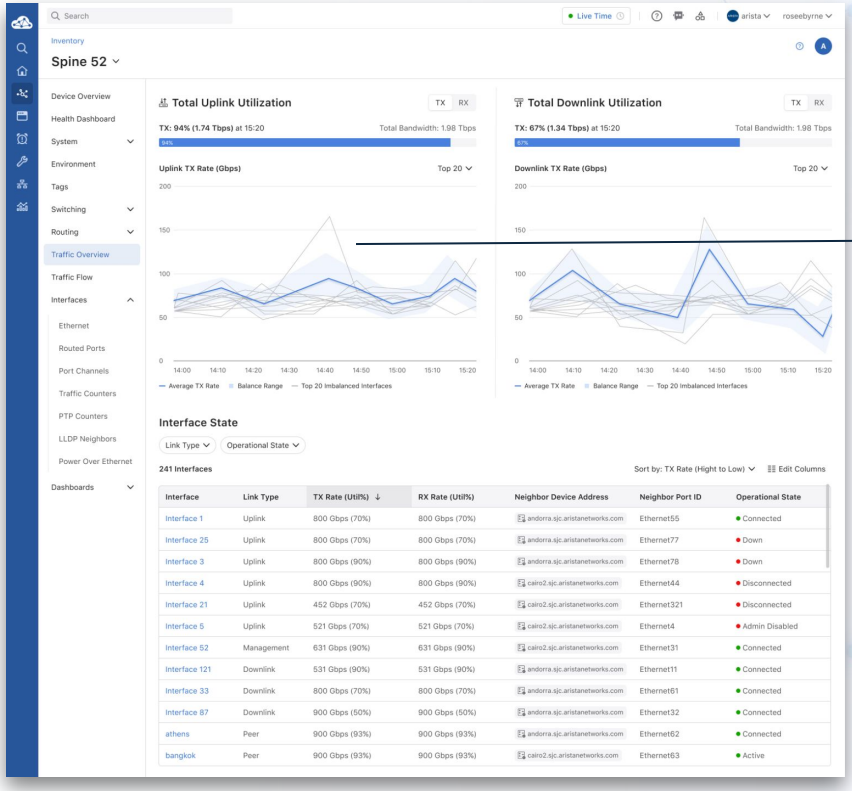
CloudVision

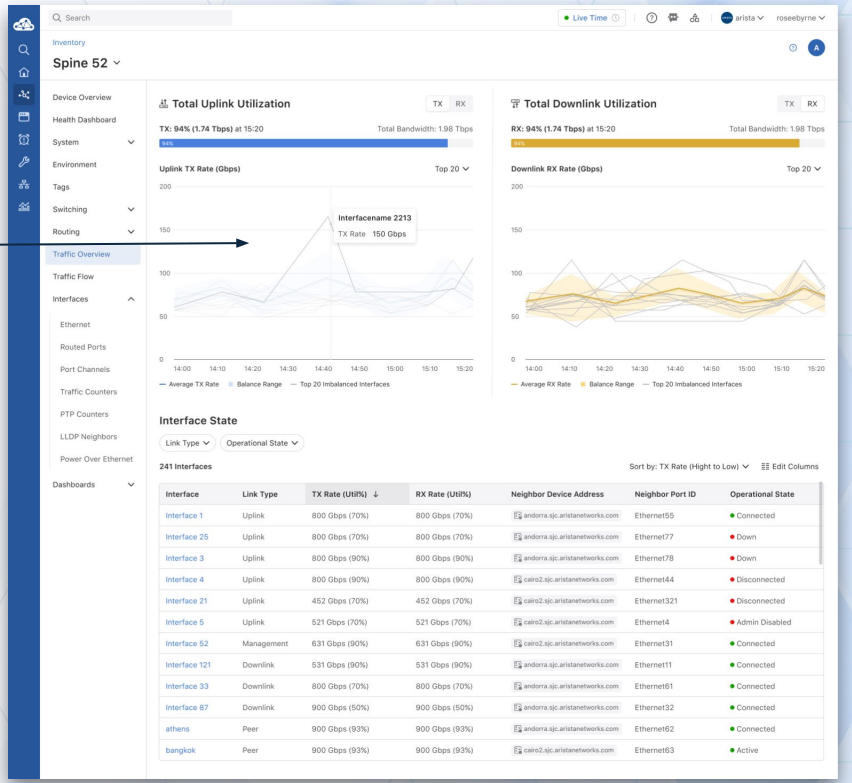


Device	Direction	Uplink Balance	Uplink Rate (Util%)	Number of Uplinks	Downlink Balance	Downlink Rate (Util%)	Number of Downlinks
Spine 52	TX	80.53 Gbps	800.22 Gbps (70.4%)	28	80 Gbps	800 Gbps (70%)	28
	RX	53.33 Gbps	1.03 Kbps (70.3%)	28	53 Gbps	800 Gbps (70%)	28
beagle	TX	40 Gbps	800 Gbps (90%)	20	40 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

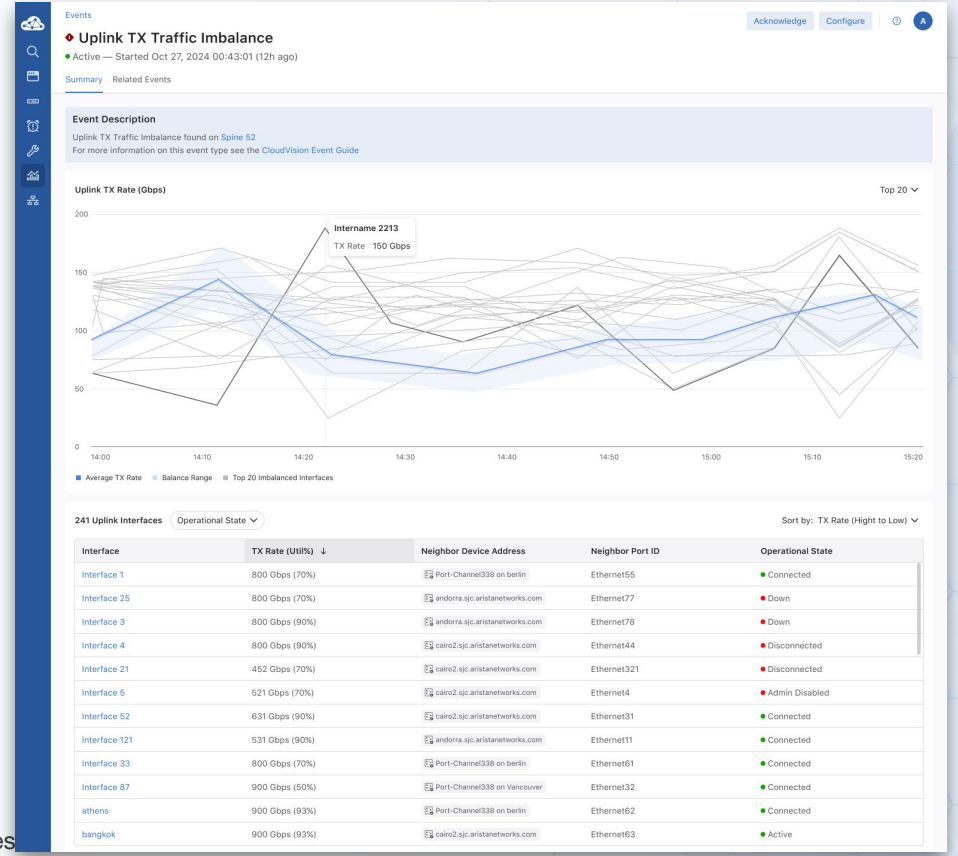
 |

ECMP Per Device Traffic Overview



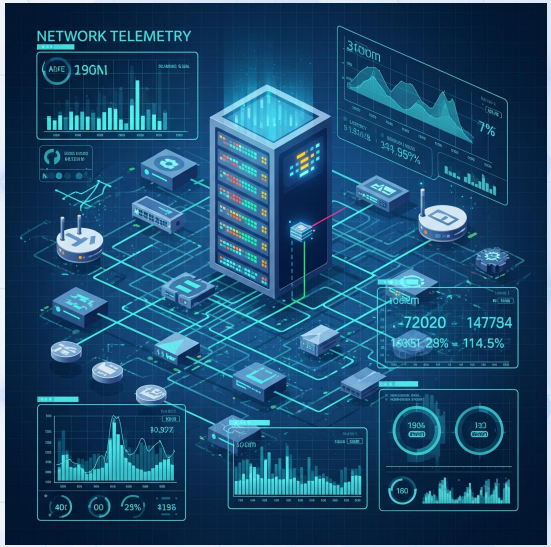


ECMP Imbalance Events



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



UNO NetDL CloudVision

Public.

ARISTA

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

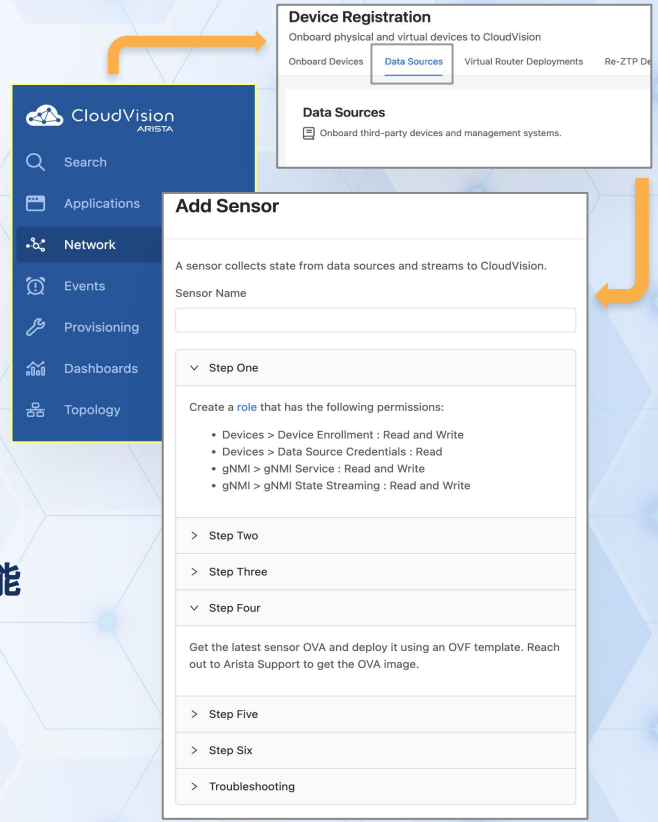
FAQ

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)

ベアメタル
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)