cisco The bridge to possible

Predictive Networks are THERE !!

JP Vasseur PhD, Cisco Fellow, VP engineering ML and Data Science
@jpvasseur, jpv@cisco.com

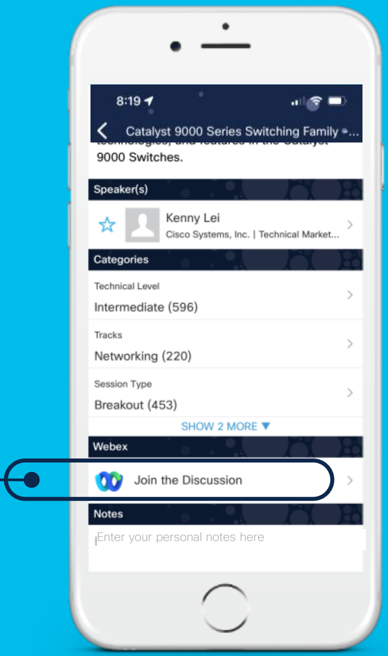
Cisco Webex App

Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space



8:19

Catalyst 9000 Series Switching Family ...

9000 Switches.

Speaker(s)

Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

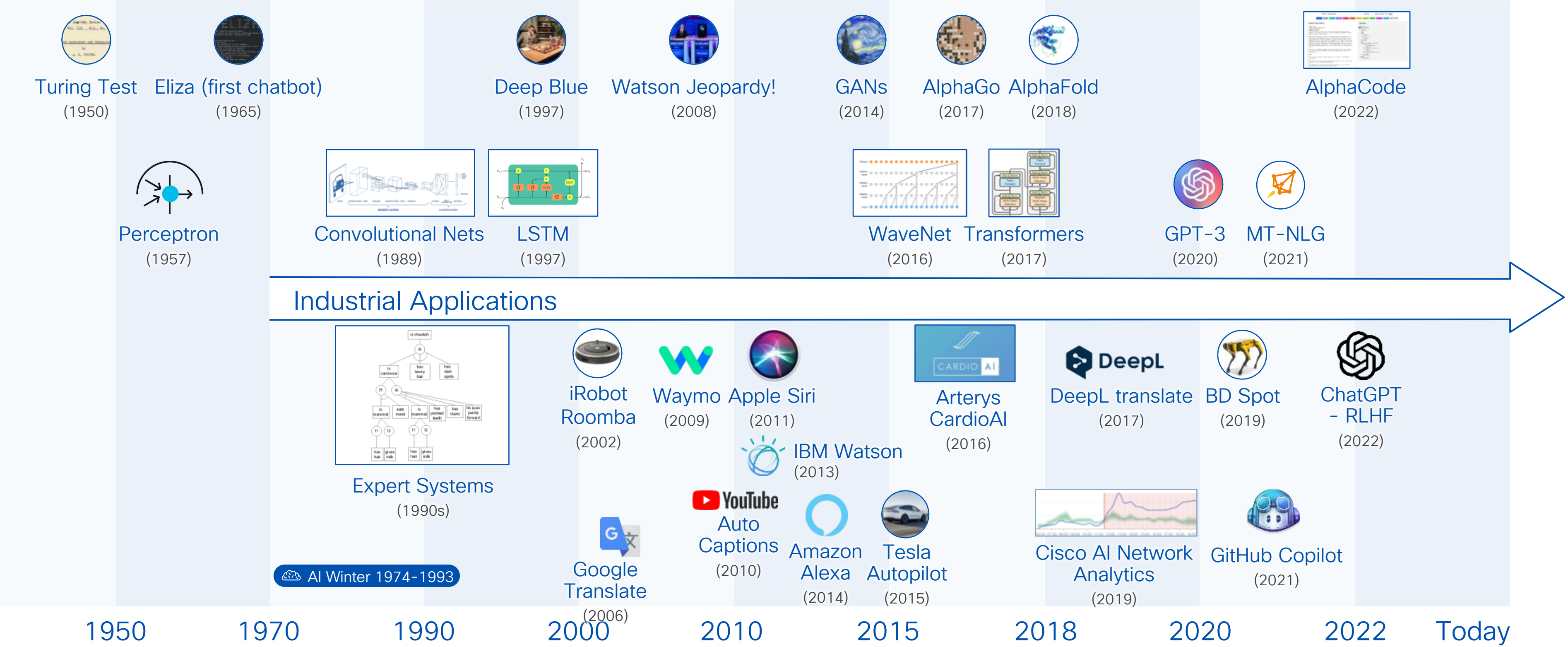
Enter your personal notes here

Webex spaces will be moderated
until February 24, 2023.

ML/AI in Networking

Predictive Networks

A brief history of AI/ML and its applications



Research & Demonstrators

Turing Test (1950)

Eliza (first chatbot) (1965)

Deep Blue (1997)

Watson Jeopardy! (2008)

GANs (2014)

AlphaGo (2017)

AlphaFold (2018)

AlphaCode (2022)

Perceptron (1957)

Convolutional Nets (1989)

LSTM (1997)

WaveNet (2016)

Transformers (2017)

GPT-3 (2020)

MT-NLG (2021)

Industrial Applications

Expert Systems (1990s)

iRobot Roomba (2002)

Waymo (2009)

Apple Siri (2011)

IBM Watson (2013)

Arterys CardioAI (2016)

DeepL translate (2017)

BD Spot (2019)

ChatGPT – RLHF (2022)

Google Translate (2006)

YouTube Auto Captions (2010)

Amazon Alexa (2014)

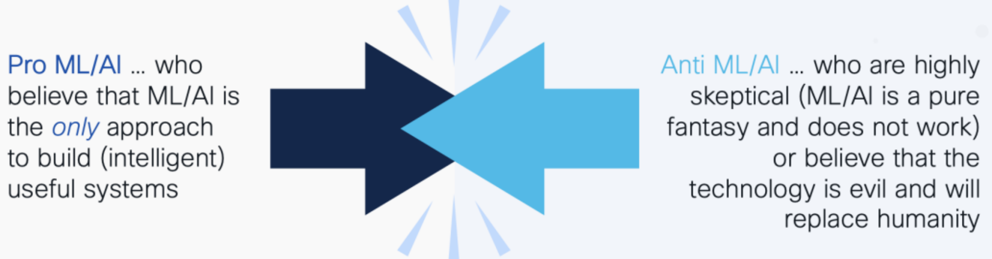
Tesla Autopilot (2015)

Cisco AI Network Analytics (2019)

GitHub Copilot (2021)

AI Winter 1974-1993

Why being skeptical about ML/AI?



Pro ML/AI … who believe that ML/AI is the *only* approach to build (intelligent) useful systems

Anti ML/AI … who are highly skeptical (ML/AI is a pure fantasy and does not work) or believe that the technology is evil and will replace humanity

- A bit of fatigue about “AI”
- Over promise, Over statements , …

At Cisco we started developing ML products a decade ago

- We have learned a lot in more than a decade of ML/AI product development
- We have tried many approaches (several failed but many worked)
- Our ML/AI products have been deployed at scale
- Results are there and AI/ML for networking moving to the next phase ….

Cisco's AI/ML Networking Journey

IoT
2012
✓

Security: Distributed Machine Learning for 0
201

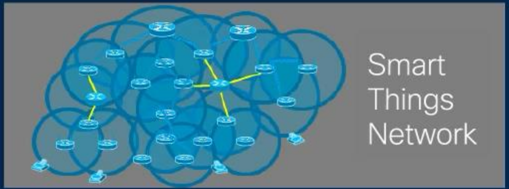
Use Case

Internet of Things

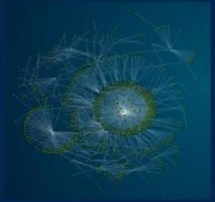
- Routing optimization
- Detection of Ddos attacks

Technology

- ML Anomaly detection
- New AI networks
- Lightweight on premise technology (40K memory)



Smart Things Network



Multi-Domain

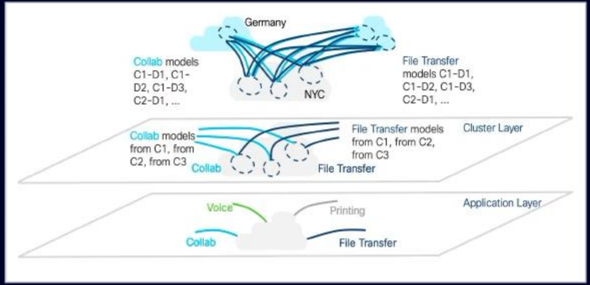
Wireless, SD-WAN,
Data Center
IoT ML Cyber Vision

Security Enterprise Networks

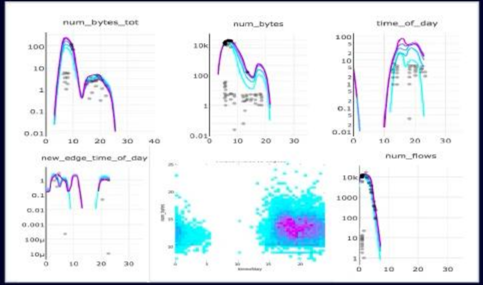
- Detection of Data ex-filtration in enterprise networks, detection of 0-day attacks
- Complex Multi-layer Security threat detection system

- On Premise ML with highly constrained environments in terms of Memory & CPU (400MBytes)
- Massively distributed ML for detection of 0-day aproach
- Multi-layer graph anomaly detection
- Anomalies related to graph and behaviors
- Use of Smart filtering thanks to user feed-back loop for anomaly filtering
- Several detection of 0-day attacks proven in the field

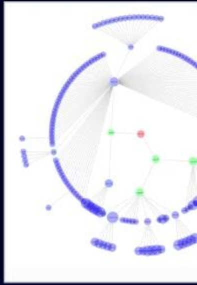
Hierarchical Graph-based anomalies


Germany
Collab models C1-D1, C1-D2, C1-D3, C2-D1, ...
NYC
File Transfer models C1-D1, C1-D2, C1-D3, C2-D1, ...
Collab models from C1, from C2, from C3
Collab
File Transfer models from C1, from C2, from C3
File Transfer
Cluster Layer
Voice
Printing
Collab
File Transfer
Application Layer

Dynamic Behavioral Models



Dynamics Device



Predictive Networks

Future

Algorithms

Demos

Predictive Networks Objectives

Results

Architecture

Predictive Networks Objectives

Imagine a world (only) reacting with no learning?

The Internet

The Internet has been Reactive for 35 years…

- Routing/QoS inherently static
- Multiple Recovery mechanisms using Protection and Restoration
  - Relies of fast detection of failure, followed by rerouting
  - Optical, Fast IGP (OSPF, IS-IS), IP FRR BGP, MPLS FRR
- Few Adaptive strategies based on recent events (backoff, …) or approximate future

No learning …

The Human Brain

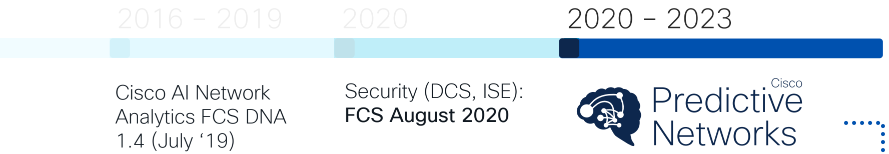
- **Learns** process not entirely known: nature versus nurture, build a model of the world (observation), ability to predict seems central, experience ***(Plasticity)***
- **Predicts** (e.g predictive coding) – Various theories
- **Plans** (higher executive functions)

Predictive Networks **(Networking "Brain")**

The Predictive Internet:

- **Builds** (ML/Statistical) models of the world (Internet & Application)
- **Predicts** potential issues (application experience)
- **Learns** and keeps improving (Telemetry)
- **Plans** with Automation

cisco Live!



2016 – 2019

Cisco AI Network Analytics FCS DNA 1.4 (July ‘19)

2020

Security (DCS, ISE):
FCS August 2020

2020 – 2023

Cisco Predictive Networks

Objectives of a Predictive Internet

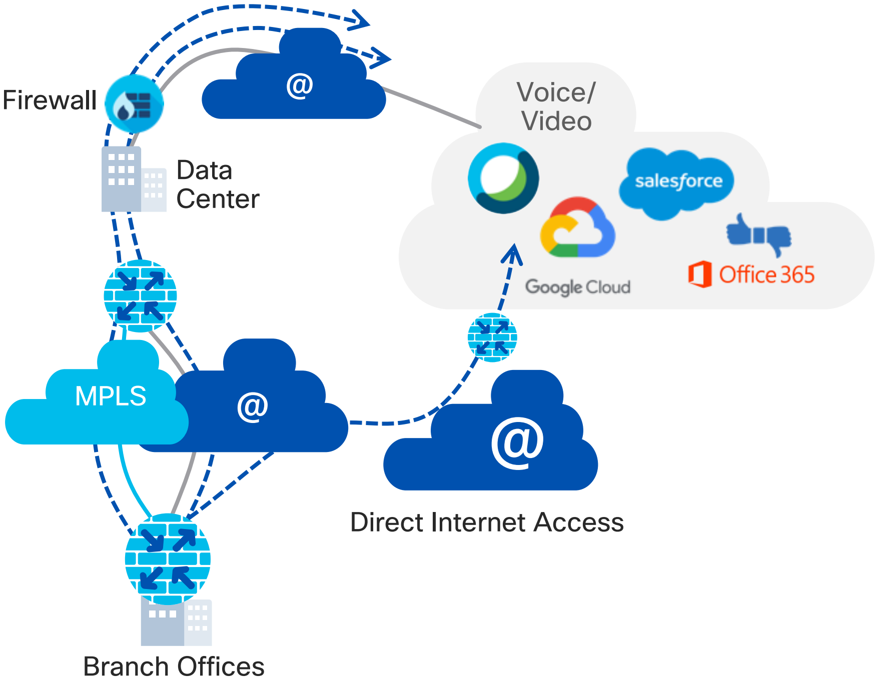
Predictive Networks is about:

Use of Predictive
(combined with Reactive)

Connectivity failures
& Application Experience

Self Healing Networks
with Trusted Automation

Short Term vs Long Term Predictions & Recommendations



Predictive Engine

Short Term Prediction (STP):

"Alto predicts Application SLA violation for Voice traffic along Internet path today from 4pm to 6pm" => Reroute to MPLS tunnels

STP uses several ML algorithms to issue "real-time" predictions

Long Term Prediction (LTP):

"Analytics shows that using the path P2 instead of P2 for O365 between the sites S1 and S2 will lead to 30% of SLA violation"

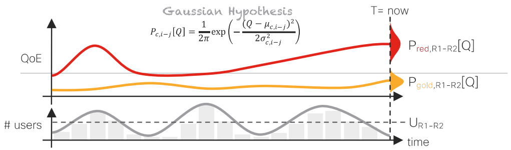
LTP loos at historical data combined with a number of metrics (stability, what-if, …) combined with prediction to make recommendation.

Real Time Prediction (RTP) is *under investigations …*

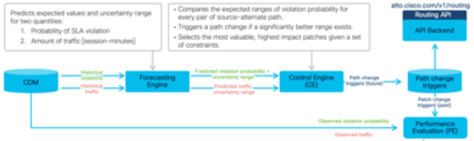
Predicting on the Internet

The notion of predicting Application failures implies that the engine predicts ***before*** it happens, in contrast with reactive approach that tries to minimizes the duration of the failure, but it is too late ☹

Our system is using various learning strategies:

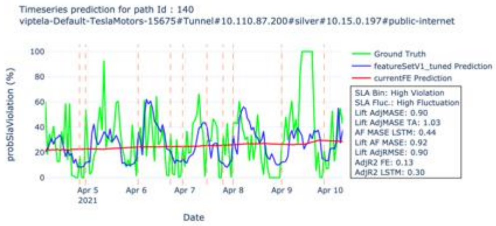


Statistical Model

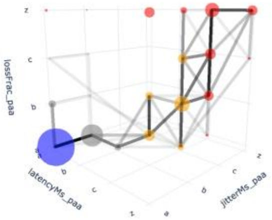


Predicts expected values and uncertainty range for two quantities:
1. Probability of SLA violation
2. Amount of traffic [session-minutes]
• Compares the expected ranges of violation probability for every pair of source-alternate path.
• Triggers a path change if a significantly better range exists.
• Selects the most valuable, highest impact patches given a set of constraints.
alto.cisco.com/v1/routing
Routing API
API Backend
CDM
Historical violations
Historical traffic
Forecasting Engine
Predicted violation probability + uncertainty range
Predicted traffic uncertainty range
Control Engine (CE)
Path change triggers (future)
Path change triggers
Patch change triggers (past)
Observed violation probability
Observed traffic
Performance Evaluation (PE)

Dynamic Model



LSTM



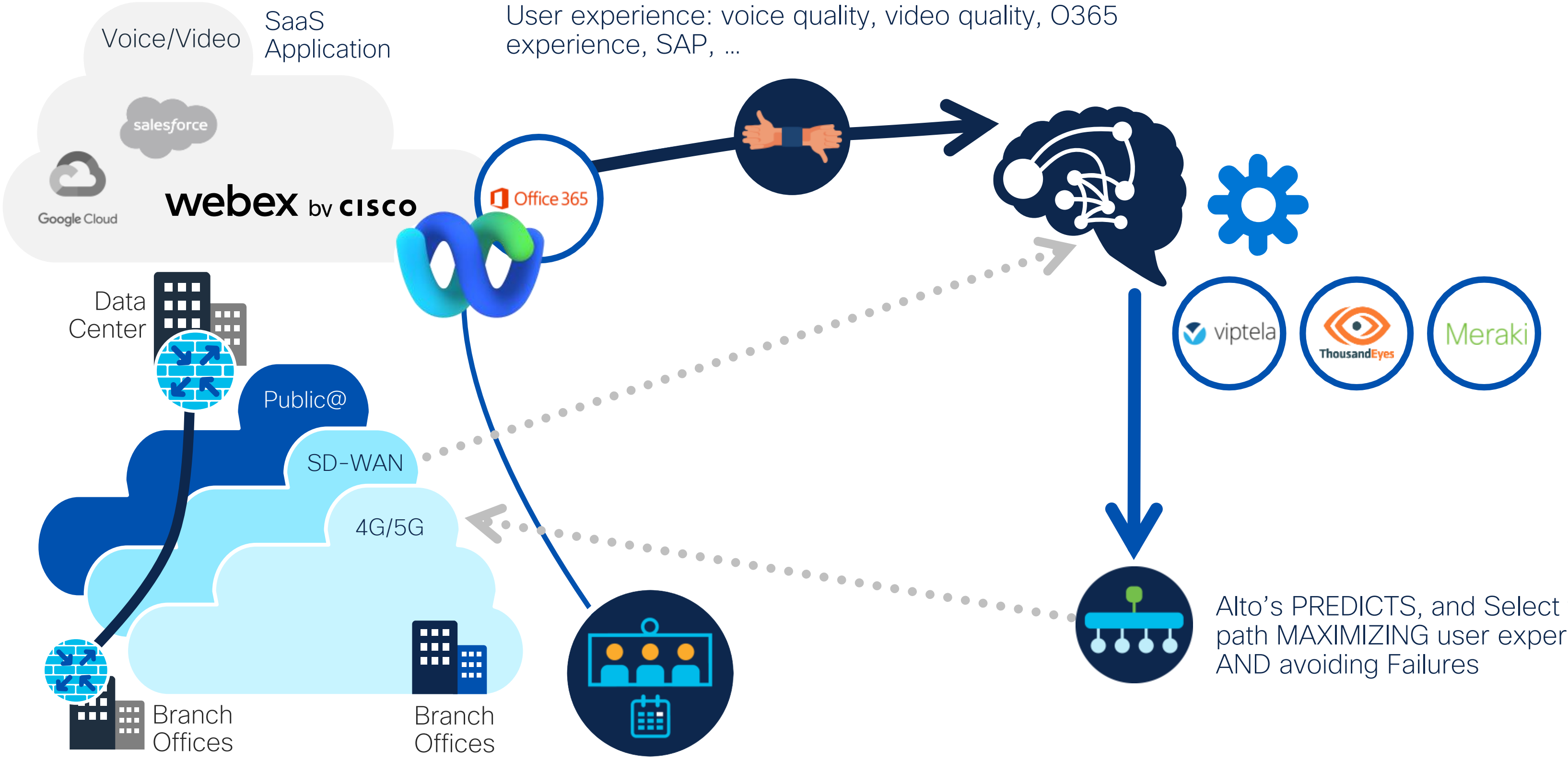
State Transition Learning

Video

Architecture

Architecture

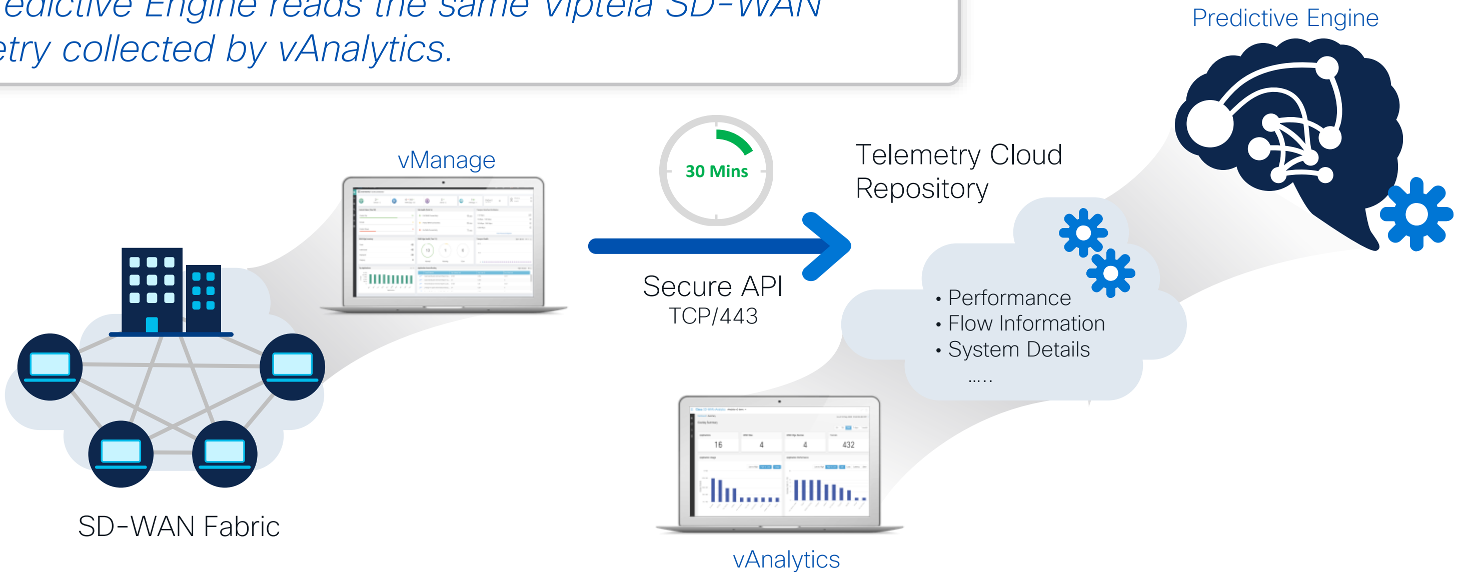
User experience: voice quality, video quality, O365 experience, SAP, …



Alto's PREDICTS, and Select the path MAXIMIZING user experience AND avoiding Failures

SD-WAN Telemetry Overview

The Predictive Engine reads the same Viptela SD-WAN telemetry collected by vAnalytics.



vManage

30 Mins

Secure API
TCP/443

Telemetry Cloud Repository

- Performance
- Flow Information
- System Details

.....

Predictive Engine

SD-WAN Fabric

vAnalytics

On-Prem or Cloud-Hosted SD-WAN (vManage)

Predictive Engine and Telemetry Repository are cloud hosted

Telemetry Details

System Details

Device Information

- Hostname
- Device Model
- System IP
- Latitude
- Longitude
- Reachability status
- Site-Id
- Org-Id

TLOC/Wan Interface

- Color
- Carrier
- Admin State
- Interface
- Preference
- Weight
- Device Name
- Private and Public IP address

Network Performance

PFR Statistics

[BDF Tunnel Probing Information]

- Latency
- Loss
- Jitter
- Hostname
- DeviceID
- Time of Day
- Local/Remote Color
- Local/Remote System IP
- Tunnel Rx/Tx Octets

CXP Statistics

[SaaS HTTP Probing Information]

- Loss
- Latency
- Application
- VPN ID
- DeviceID
- Interface
- Gateway System IP
- Best Path
- Local/Remote Color
- MSFT Quality Labels

Network Usage

DPI Statistics

- Application/Family
- Local/Remote IP Address
- Local/Remote Port
- Ingress Local/Remote Color
- Egress Local/Remote Color
- Local/Remote System IP
- Egress Interface
- Time of Day
- VPN ID
- Packets
- Octets

Interface Statistics

- VPN ID
- Rx/Tx Packets
- Rx/Tx Octets
- Rx/Tx Errors
- Rx/Tx Drops
- Rx/Tx PPS
- Platform Type
- Operational Status

Notes:

- For a medium customer we expect to process around: 1GB of PFR and 30GB of DPI of data per day.
- Telemetry may arrive in slightly different formats based on software version, source platform (cEdge, vEdge) and is reconciled during ingestion.

Results

Results

Long Term Predictions

SLA Violations Across the World

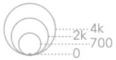
and how much Predictive Networks can help

0
CUSTOMERS

0
COUNTRIES

2
CITIES

Cisco
redictive
etworks

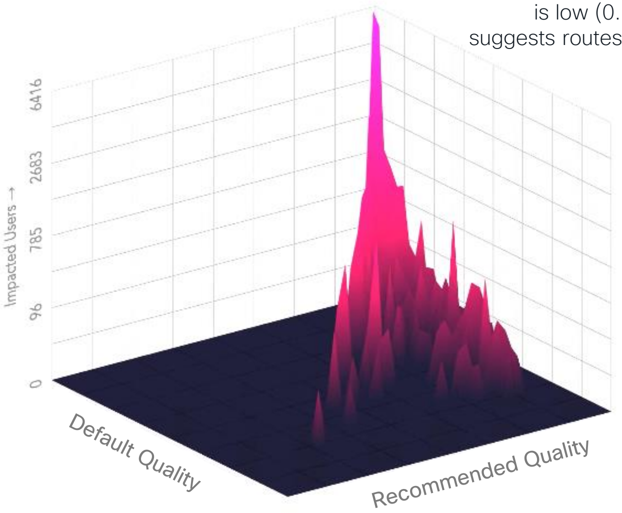


Quality metrics vs. Num users & sites

All Apps, Impacted Users

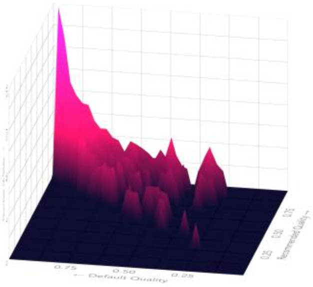
Tallest peak: Large number of users where Alto further improves the quality of an already good quality route

High density of peaks where reco quality significantly better than default quality: Many Impacted Users at sites whose default quality is low (0.25 to 0.50), but Alto suggests routes with recommended quality (0.50 to 1.00)

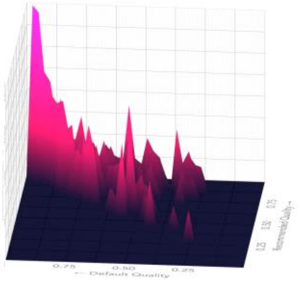


Recommendation Quality > Default Quality without Alto: Alto’s recommended quality is significantly greater than default quality. No peaks where default quality > reco quality

All apps, Sites



All Apps, Impacted Users



High density of peaks where recommended quality significantly better than default quality: Several sites where recommendation quality is superior to default quality

Tallest peak: Many sites where default and recommended quality is close to 1

Recommendation Quality > Default Quality without Alto
Several sites and users where default quality < 0.5 and recommended quality is close to 1

Legend

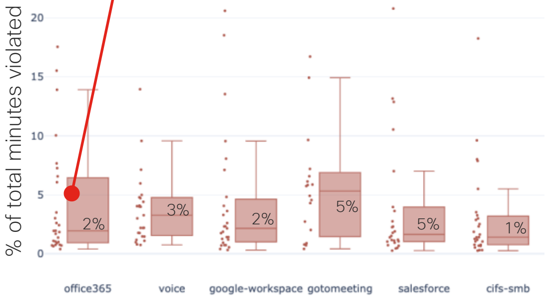
- X : Efficacy; Y: Default Quality
- Z : Number of Impacted Users (left) and Sites (right)
- Z-Scale: Scaled with cubic root
- Color: Colored by the value of Z-axis
- Selected data: all sites with saved users >= 1
- 40 x 40 bins on XY plane

Short-term Prediction: Summary metrics

Data: 6 months (June-Nov 2021); 27 customers that are diverse.

Percentage of violated minutes

Around 2.% of total session minutes observed are violated for “office365”

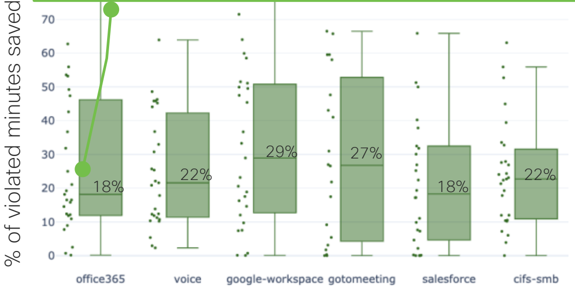


“office365” has 2% & “voice” has the 3% of total session minutes that violate SLA thresholds.

- X = App; Y=Violated session minutes / total session minutes
- Each dot = 1 customer

Percentage of violated minutes saved

At median, 18% of violated session minutes are saved for “office365” across customers. A 75th percentile of 46% of violated minutes are saved.

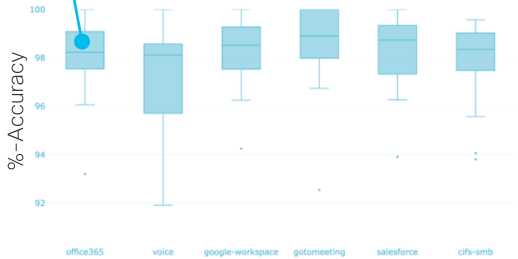


- At median levels, **18% to 27%** of violated minutes are saved across customers
- At 75th percentiles, **32% to 52%** of violated minutes are saved

- X = App; Y= %-age of violated session minutes saved
- Each dot = 1 customer

Accuracy

“office365” has median, 25thp and 75th p accuracy ~98%,

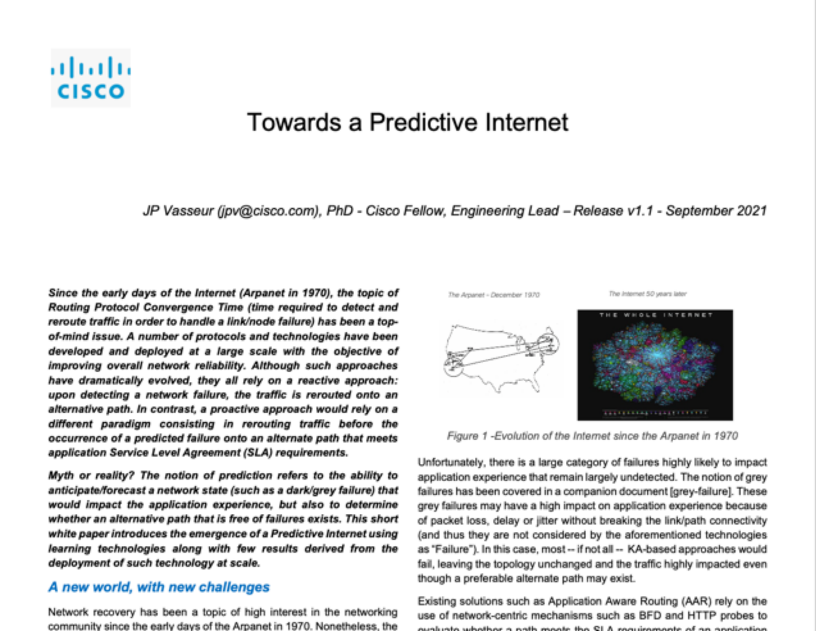


Alto has high **~98%+ median accuracy for apps.**

- X = App;
- Y= Accuracy = Num Recommendations with positive savings / Total Num Recommendations
- Each dot = 1 customer

Demos

Want to know more about the Predictive Networks?



Towards a Predictive Internet

JP Vasseur (jpv@cisco.com), PhD - Cisco Fellow, Engineering Lead – Release v1.1 - September 2021

Since the early days of the Internet (Arpanet in 1970), the topic of Routing Protocol Convergence Time (time required to detect and reroute traffic in order to handle a link/node failure) has been a top-of-mind issue. A number of protocols and technologies have been developed and deployed at a large scale with the objective of improving overall network reliability. Although such approaches have dramatically evolved, they all rely on a reactive approach: upon detecting a network failure, the traffic is rerouted onto an alternative path. In contrast, a proactive approach would rely on a different paradigm consisting in rerouting traffic before the occurrence of a predicted failure onto an alternate path that meets application Service Level Agreement (SLA) requirements.

Myth or reality? The notion of prediction refers to the ability to anticipate/forecast a network state (such as a dark/grey failure) that would impact the application experience, but also to determine whether an alternative path that is free of failures exists. This short white paper introduces the emergence of a Predictive Internet using learning technologies along with few results derived from the deployment of such technology at scale.

A new world, with new challenges

Network recovery has been a topic of high interest in the networking community since the early days of the Arpanet in 1970. Nonetheless, the

The Arpanet - December 1970

The Internet 50 years later

Figure 1 -Evolution of the Internet since the Arpanet in 1970

Unfortunately, there is a large category of failures highly likely to impact application experience that remain largely undetected. The notion of grey failures has been covered in a companion document [grey-failure]. These grey failures may have a high impact on application experience because of packet loss, delay or jitter without breaking the link/path connectivity (and thus they are not considered by the aforementioned technologies as "Failure"). In this case, most -- if not all -- KA-based approaches would fail, leaving the topology unchanged and the traffic highly impacted even though a preferable alternate path may exist.

Existing solutions such as Application Aware Routing (AAR) rely on the use of network-centric mechanisms such as BFD and HTTP probes to

From Dark to Grey Failures in the Internet

JP Vasseur (jpv@cisco.com), PhD, Cisco Fellow, Engineering Lead - June 2021

Quantifying Network Path Dynamics using Time-series Features

Sambarta Dasgupta (sambdasg@cisco.com) – Lead Data Scientist, Vinay Kolar (vinkolar@cisco.com) – Principal Engineer, JP Vasseur (jpv@cisco.com), PhD - Cisco Fellow Engineering Lead -January 2022

Application-aware Oscillation Detection in Internet Paths

Vinay Kolar (vinkolar@cisco.com) – Principal Engineer, JP Vasseur (jpv@cisco.com), PhD - Cisco Fellow Engineering

Large-scale Internet Path modelling and applications

Vinay Kolar, Ph.D., Principal Engineer ,JP Vasseur, Ph.D., Cisco Fellow, , Mukund Raghuprasad (myelahan@cisco.com) - Sept 2020

1 Introduction

2 Internet-Scale Data collection

An Analysis of Network KPI variability for various SaaS application in the Internet

JP Vasseur (jpv@cisco.com), Ph.D., Cisco Fellow, Mukund Yelahanka Raghuprasad (myelahan@cisco.com), Vinay Kolar (vinkolar@cisco.com), Ph.D.

October 2021

Predicting which kind of failures?

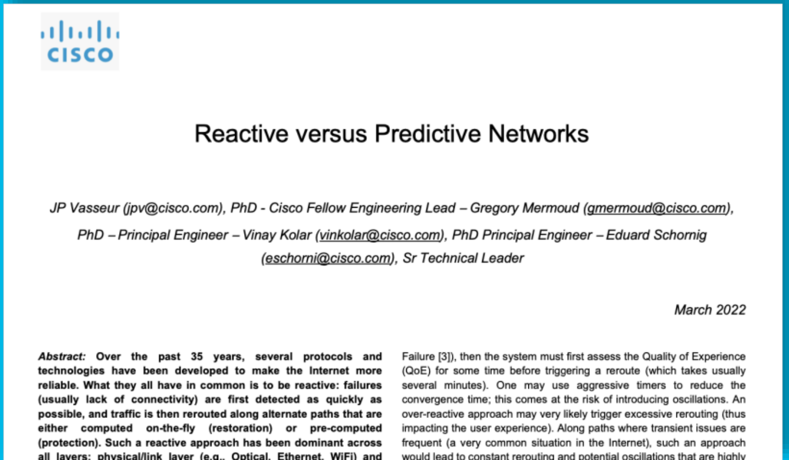
Core focus has been on Dark Failures (lack of connectivity) followed by "fast" traffic rerouting: This is (almost) SOLVED

- *Fast failure detection (multilayer, KA BFD, ISIS/OSPF fast hellos)*
- *Protection (Optical, MPLS TE FRR, BGP, …) & Fast restoration (iSPF, …)*

Predictive vs Reactive

There is Predictive versus Reactive …
they are complementary ….

Predictive allows for **AVOIDING issues with high accuracy**; for all non predictable/predicted issues, reactive will react



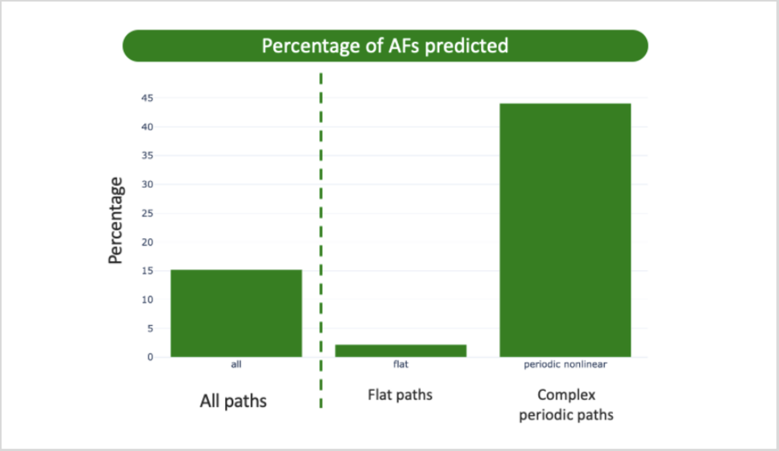
Reactive versus Predictive Networks

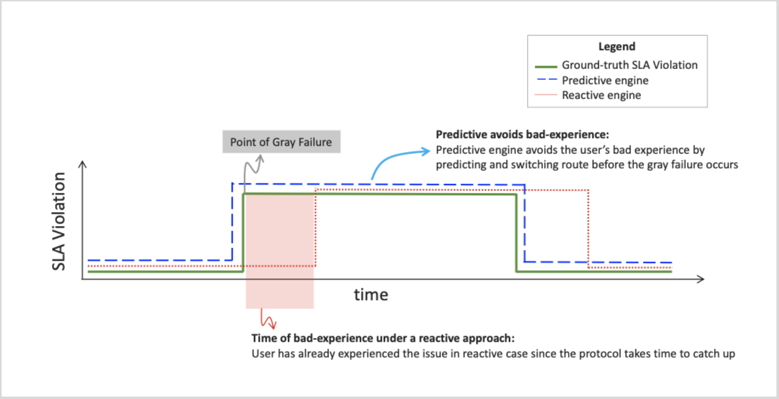
JP Vasseur (jpv@cisco.com), PhD - Cisco Fellow Engineering Lead – Gregory Mermoud (gmermoud@cisco.com), PhD – Principal Engineer – Vinay Kolar (vinkolar@cisco.com), PhD Principal Engineer – Eduard Schornig (eschorni@cisco.com), Sr Technical Leader

March 2022

Abstract: **Over the past 35 years, several protocols and technologies have been developed to make the Internet more reliable. What they all have in common is to be reactive: failures (usually lack of connectivity) are first detected as quickly as possible, and traffic is then rerouted along alternate paths that are either computed on-the-fly (restoration) or pre-computed (protection). Such a reactive approach has been dominant across all layers: physical/link layer (e.g., Optical, Ethernet, WiFi) and**

Failure [3]), then the system must first assess the Quality of Experience (QoE) for some time before triggering a reroute (which takes usually several minutes). One may use aggressive timers to reduce the convergence time; this comes at the risk of introducing oscillations. An over-reactive approach may very likely trigger excessive rerouting (thus impacting the user experience). Along paths where transient issues are frequent (a very common situation in the Internet), such an approach would lead to constant rerouting and potential oscillations that are highly

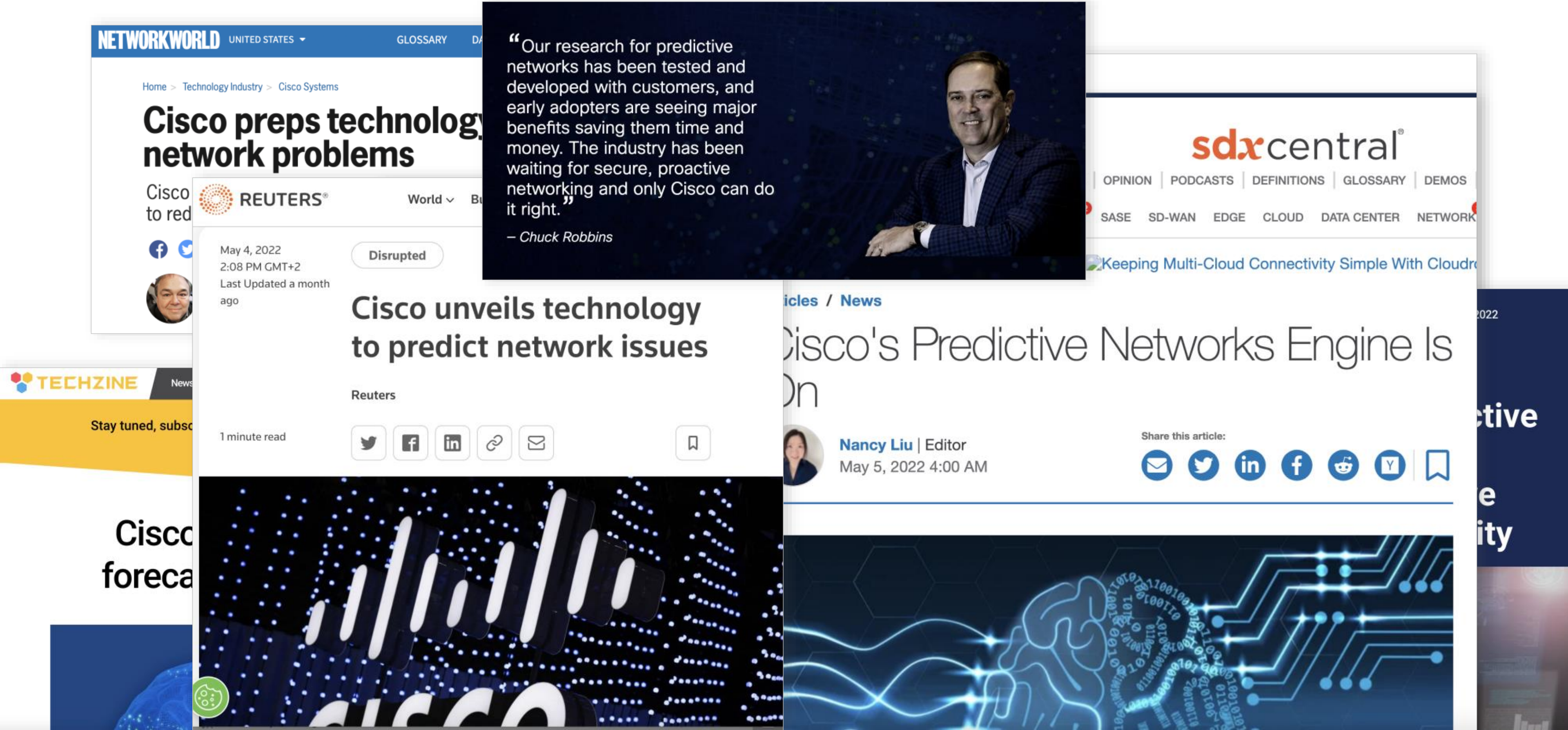




Predictive avoids bad-experience:
Predictive engine avoids the user's bad experience by predicting and switching route before the gray failure occurs

Time of bad-experience under a reactive approach:
User has already experienced the issue in reactive case since the protocol takes time to catch up

Announcing our vision for Predictive Networks



NETWORKWORLD UNITED STATES ▾ GLOSSARY DA

Home > Technology Industry > Cisco Systems

Cisco preps technolog
network problems

Cisco
to red

> “Our research for predictive networks has been tested and developed with customers, and early adopters are seeing major benefits saving them time and money. The industry has been waiting for secure, proactive networking and only Cisco can do it right.”
>
> *– Chuck Robbins*

REUTERS® World ⌄ Bu

May 4, 2022
2:08 PM GMT+2
Last Updated a month ago

Disrupted

Cisco unveils technology to predict network issues

Reuters

1 minute read

TECHZINE News

Stay tuned, subsc

Cisco
foreca

sdxcentral®

OPINION | PODCASTS | DEFINITIONS | GLOSSARY | DEMOS

SASE SD-WAN EDGE CLOUD DATA CENTER NETWORK

Keeping Multi-Cloud Connectivity Simple With Cloudr

icles / News

Cisco's Predictive Networks Engine Is On

Nancy Liu | Editor
May 5, 2022 4:00 AM

Share this article:

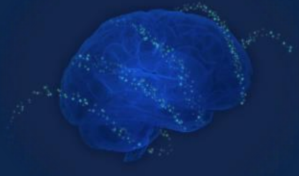
022

ctive
e
ity

Predictive Networks First milestones shipping with ThousandEyes and vAnalytics

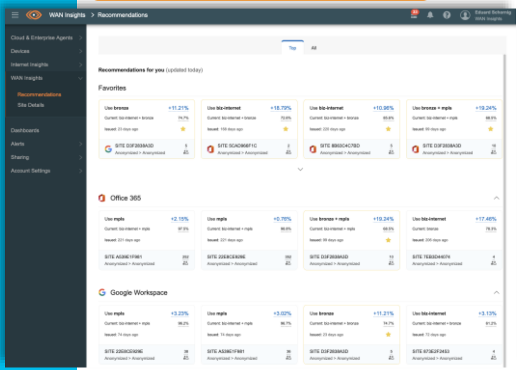
Press Release

New Cisco Technology Can Predict Network Issues Before They Happen

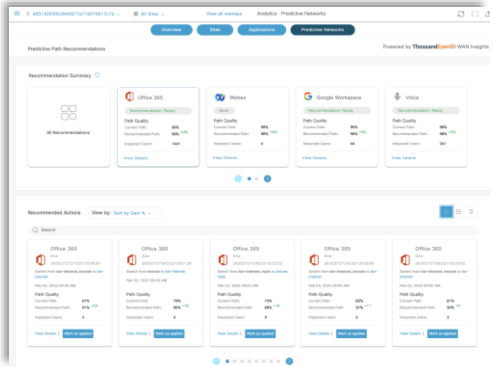


> “Our research for predictive networks has been tested and developed with customers, and early adopters are seeing major benefits saving them time and money. The industry has been waiting for secure, proactive networking and only Cisco can do it right.”
>
> – *Chuck Robbins*

Thousand Eyes (WAN Insights)



vAnalytics (Predictive Networks)



CISCO Live!

Future

- Predictive Networks applies to a number of Networking areas …

Predictive SASE

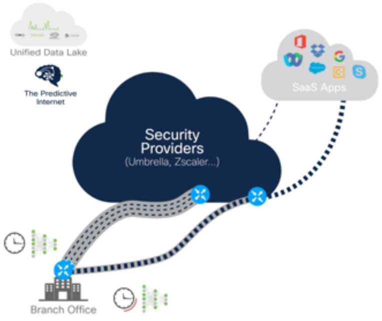
Customer Outcome: existing solution sent traffic to the "closest" CSP PoP with no SLA guarantees. The solution would **learn** and **predict** which PoP to select, which path to use and which traffic to send via the SIG tunnel. Ability to combine Security and Guaranteed application SLA in a very dynamic environment. Application experience feedback used for path selection (first time).

Technology: Central learning engine (Alto) with new algorithms, full automation (possible with Viptela) on tunnel to setup and policy to use. Viptela + Meraki Frontizo (with some effort).

Risk: Moderate, moderate engineering work.

Time-frame: (with cross-BU collaboration) 12 months

Differentiation: High with zScaler, PAN, GCP, …



Predictive Hybrid

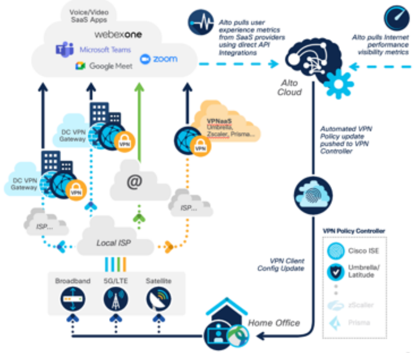
Customer Outcome: learn and predict which traffic to send to VPN tunnel, which VPN tunnel to build, which interface to use AT HOME to guarantee best user experience.

Technology: Central learning engine (Alto) with new algorithms, full automation (via controller like ISE, …), diverse telemetry (application, local engine LAN/Netflow). **First totally autonomous agent for Hybrid (could be embarked on Laptop, smart phone)**

Risk: High (full autonomous, algo, number of dependencies, results)

Time-frame: (with cross-BU collaboration) 18 months

Differentiation: High with zScaler, Versa, PAN, …



Predictive AppD

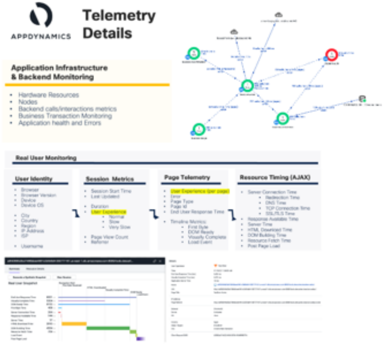
Customer Outcome: learn and predict Application issues/anomalies using AppD Telemetry for large home grown applications (today: anomaly detection + root causing). **"killer-app" Predictive for call center and corporate remote users for SAP.**

Technology: Central learning engine (Alto) with new algorithms, custom-based AppD telemetry, with potential automation for mobile apps, remote sites, + may application hosting in DC

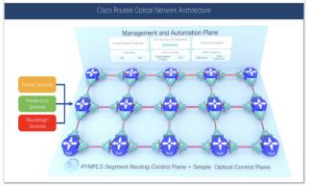
Risk: Moderate (new telemetry, app dependency)

Time-frame: (with cross-BU collaboration) 12 months

Differentiation: High with DataDog (no doing Predictive)


APPDYNAMICS
Telemetry Details
Application Infrastructure & Backend Monitoring
Hardware Resources
Nodes
Backend calls/interactions metrics
Business Transaction Monitoring
Application health and Errors
Real User Monitoring
User Identity
Session Metrics
Page Telemetry
Resource Timing (AJAX)

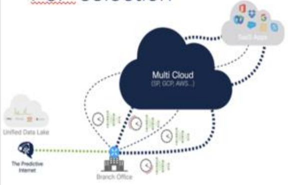
SP Use Case 1
Predictive Routed Optical Networks


Cisco Routed Optical Network Architecture
Management and Automation Plane
IP/MPLS Segment Routing Control Plane + Simple Optical Control Plane

SP Use Case 2
Extending Reach with Predictive SLA

ISP 1
ISP 2
ISP 3
ISP 3
ISP 4
ISP 6
ISP 1
Predictive Internet Path

SP/Hyperscaler Use Case 3
Predictive best GCP PoP selection



A Sneak Peak at Cognitive Networks

Video

Cognitive Networks in 1’

Complete your Session Survey

- Please complete your session survey after each session. Your feedback is important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (open from Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Session Catalog and clicking the "Attendee Dashboard" at https://www.ciscolive.com/emea/learn/sessions/session-catalog.html

Continue Your Education

Visit the Cisco Showcase for related demos.

Book your one-on-one Meet the Engineer meeting.

Attend any of the related sessions at the DevNet, Capture the Flag, and Walk-in Labs zones.

Visit the On-Demand Library for more sessions at ciscolive.com/on-demand.

The bridge to possible

Thank you

cisco Live!
ALL IN