Intelligent Network Infrastructure

Prepared by NGIP laboratory and Presented by Peter Chun

Peter Chun, Ph.D., P.Eng.
Manager of Technical Planning & Collaboration
Next Generation Computing Lab, Canada Research Centre

- Taught at Ryerson university where he was a faculty member teaching for advanced electronics, microprocessor systems and electric machines and electronics.
- Worked on Advanced Design Technology group at Nortel working on the first 40G SONET FEC and Ethernet Metro optical systems,
- Moved on the aerospace industry (at MDA) designing electronic systems such as next generation Canadarm, RADARSAT constellation mission system
- Expanded his horizon into mass-media (at Nevion) & consumer electronic market (at RAMBUS) where he was involved in developing and managing real-time video transport stream systems and mobile image sensor processing units.
- Engaging in technical planning and collaboration for heterogeneous device/pipe/cloud solutions for emerging 5G/MEC applications in ICT convergence environment at Huawei.

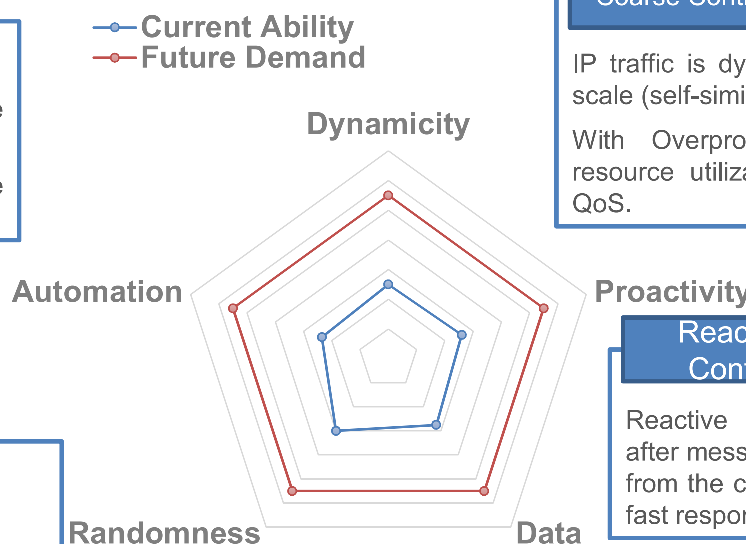
Content

Some Open Issues in IP Networks

Limited Automation

New services (e.g. IoT, 5G) exceed the manual capacity.

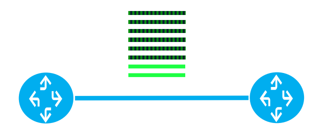
Better **automation** is required for the future networks.



Coarse Control

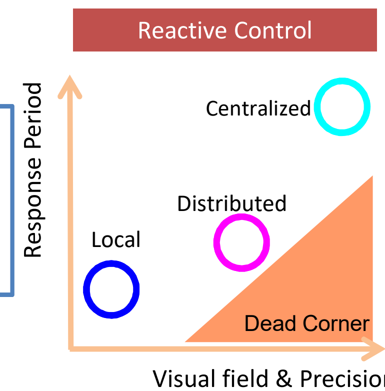
IP traffic is dynamic and bursty in any time scale (self-similarity).

With Overprovisioning or “light loading”, resource utilization is traded for barely OK QoS.



Reactive Control

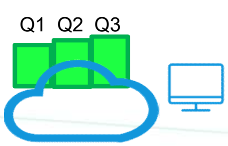
Reactive control responses to events only after messages arrive. There is a dead corner from the conflict of necessary visual field and fast response.



Poor treatment to randomness

Randomness in networks status is reality.

Just counting the average and maximum values and overprovisioning with a wide and static safe margin, is inefficient and human experience depending.



Weak Data Collection

Current data measurement and collection are auxiliary and weak， just matching demands from limited status monitoring. Future data-based control & management will require huge amount, large-scale and high frequency data collection. It needs to evolve to a well-designed, feature-rich sub-system.

Deep Learning Algorithms Applications to IP Networks

Algorithm	Description	Function	Popular Applications	App Examples in Networks Industry
DNN: Deep Neural Network	Multiple layer feed-forward neural network. Be able to fit any function Infinitely by learning.	Automatic feature extraction and mapping to complex data set	Classification, object recognition, regression	**Can be widely used. e.g. traffic prediction, failure prediction** ✓
RNN: Recurrent Neural Network	Neural network with loops. Be able to learn to remember temporal events selectively.	Automatic feature extraction from temporal data set with arbitrary length.	Temporal data processing, e.g. text understanding, text translation, speech recognition	**Can be widely used. e.g. traffic prediction, temporal optimization** ✓
CNN: Convolution Neural Network	Scan data with a set of shared localized features and learn those features	Extract common localized features efficiently. Object recognition with space invariance.	Image recognition, voice recognition	**Graph-based CNN for feature analysis on network topology. e.g. QoE analysis/prediction**
GAN: Generative Adversarial Network	Generate accurate samples with identification module using true/false signal	For problem with known positive sample set and unknown loss function.	High dimensional data generation with stochastic. e.g. image generation, processing, denoising	**Traffic sample generation, failure sample generation**
VAE: Variational Auto-Encoder	Variational inference coding in form of auto-encoder	Unsupervised learning with stochastic modeling & generation	Image generation, processing, denoising	**QoS modeling, failure modeling, traffic pattern analysis**
DRL: Deep Reinforcement Learning	DNN + Reinforcement Learning	Close loop control under complex environment, especially with high dimensional observation input	Automatic control system, robotics, gaming	**Temporal control/optimization like TE, traffic balance, topology planning, energy saving** ✓

5 Levels to Intelligent Network Infrastructure

Best Effort.
No traffic optimization

Level 0

No optimization

Examples:
- TE
- SD-WAN

Deterministic/static policy to optimize traffic.

Level 1

Traditional optimization

Examples:
- Traffic prediction
- Elephant flow identification
- QoS Inference
- Selective measurement

Open Loop Control
Implement traffic monitoring and recognition.
Manually make the policies and adjust the network configuration in period.

Level 2

ML-aided open loop control optimization

Examples:
- Intl overlay routing
- Intl traffic balance

Close Loop Control
Part of the network policies can be deployed and executed automatically and others retain human decision-making.

Level 3

Close loop control optimization by ML

All the network policies can be deployed and executed automatically. Meanwhile, retain the manually control capacity.

Level 4

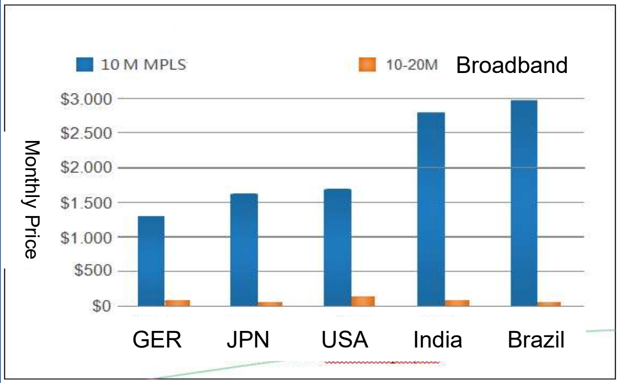
Fully automated close loop control optimization by ML

Content

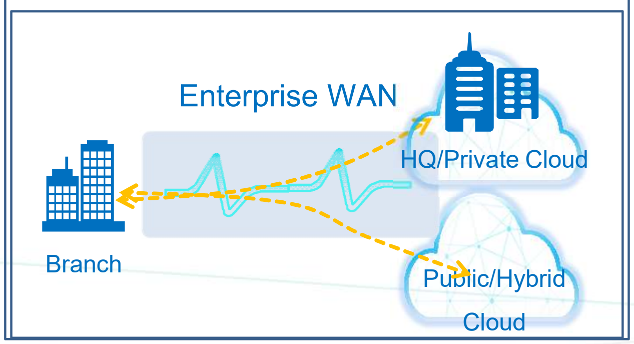
Case Study 1: Intelligent Overlay Routing (Open Issues)

Key Issues of WAN

✓ **MPLS IP VPN vs Internet: 15~50x in price**

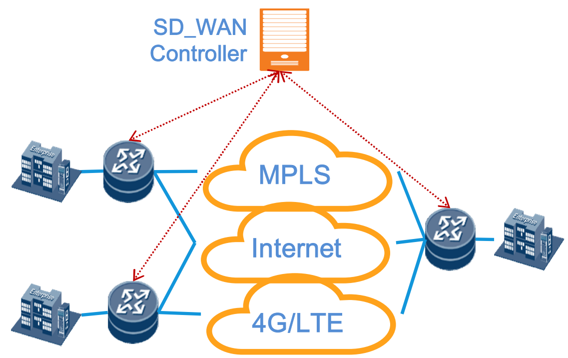


Traffic pattern is complex & constantly changing
Manual config is static, vulnerable to mistakes

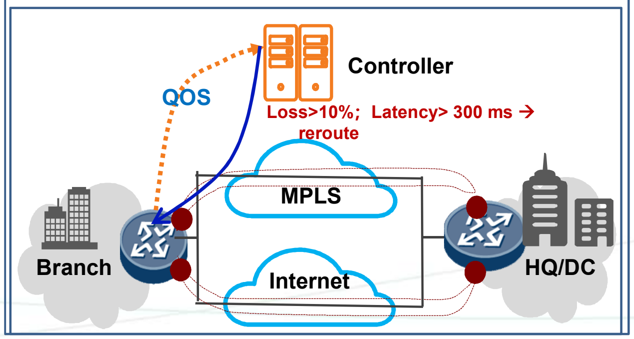


SD-WAN

SD-WAN

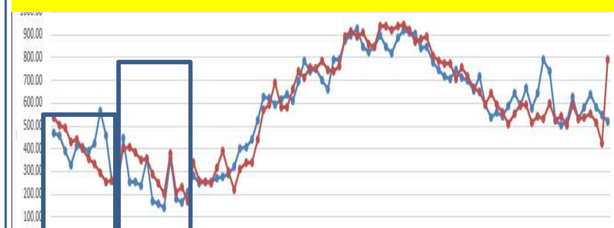


Support hybrid networking (MPLS, Internet, LTE)
Support threshold-based dynamic routing



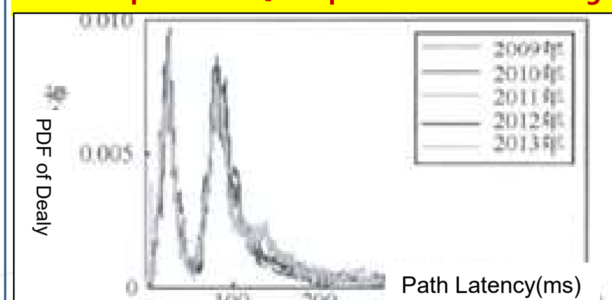
Issues to be Resolved

✓ **Threshold-based routing needs human experience to set threshold with large safe margin**



Similar daily traffic pattern can be utilized (15-min mean traffic of a day)

✓ **No proactive QoS Optimization in Routing**



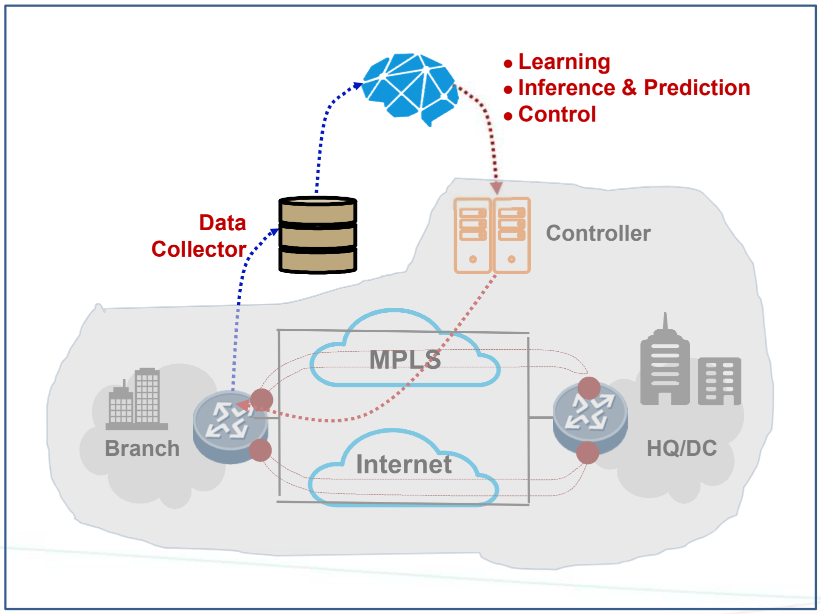
Stable Patterns in Internet latency can be utilized

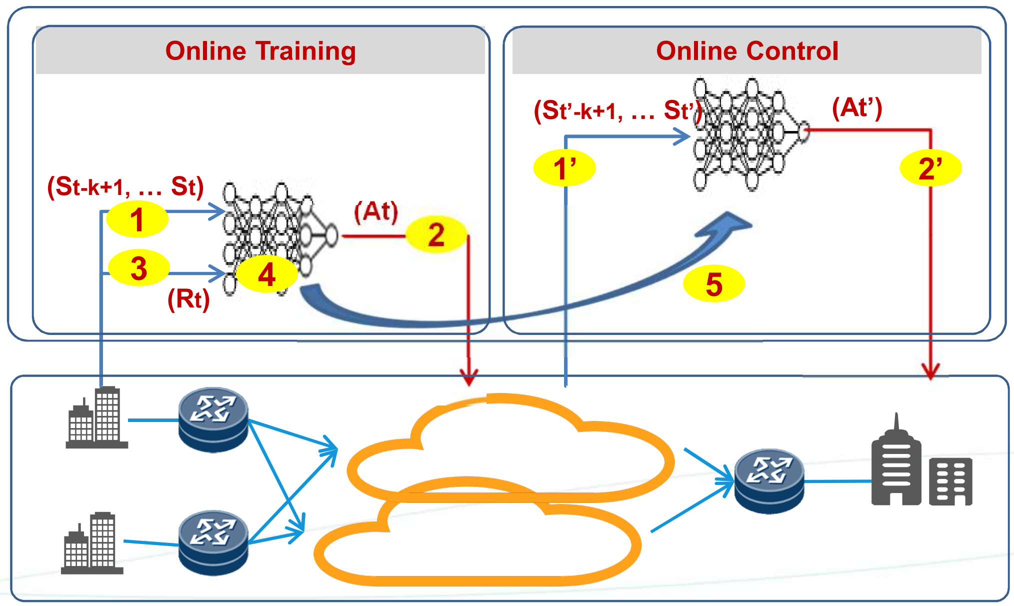
Case Study 1: Intelligent Overlay Routing (Solution)

Deep Reinforcement Learning

- ✓ Deep Learning extracts VPN traffic pattern and underlay network QoS pattern implicitly
- ✓ Reinforcement Learning controls outgoing routing at WAN gateway

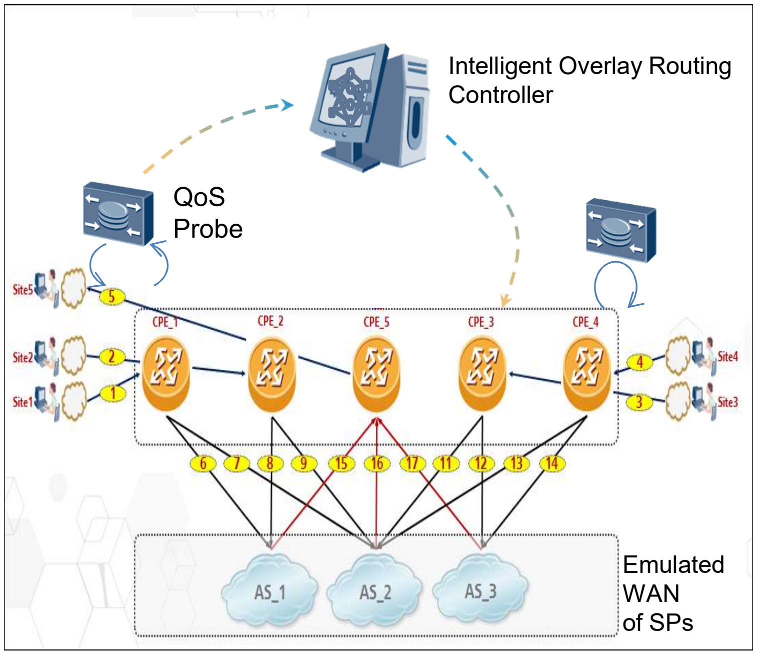
1. Collect traffic status(current & historical)
2. DRL generates routing actions and instructs routers
3. Measures QoS. Computes rewards with multiple objectives: latency, jitter, dropping, expenses
4. Improve actions for maximum long-term accumulated rewards
5. Control SD-WAN with trained DRL





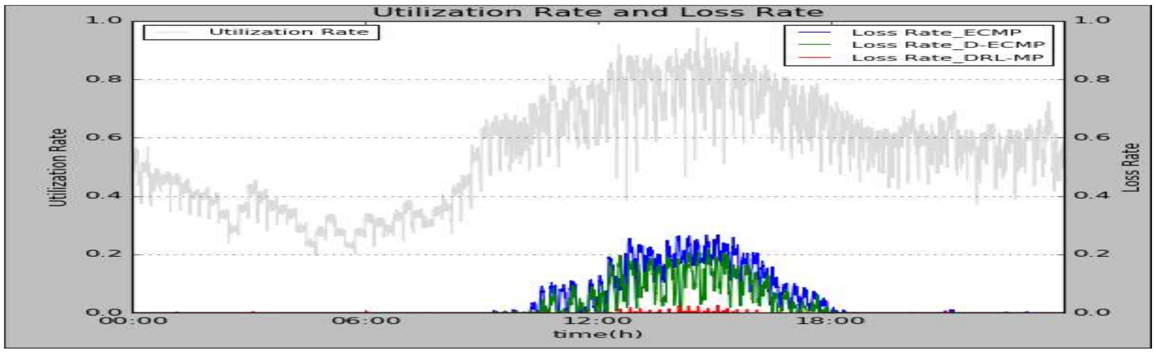
Case Study 1: Intelligent Overlay Routing (Demo Testing)

DEMO Topology

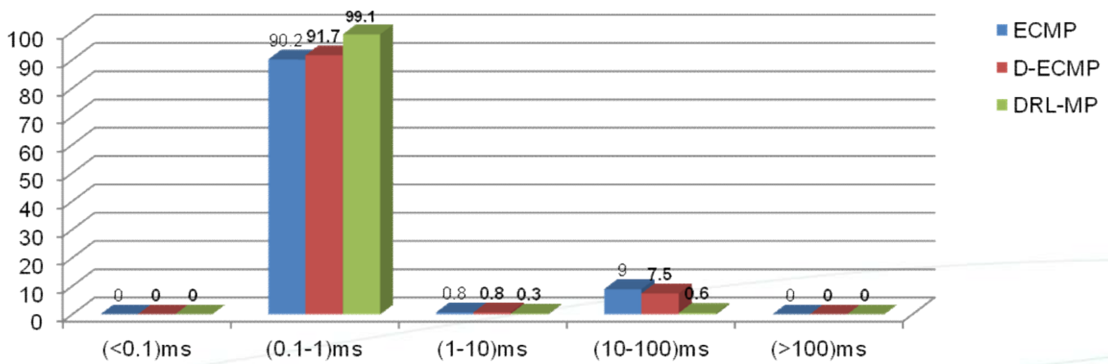


Test bed with NE40 routers
RTT measured by NEU200 probes

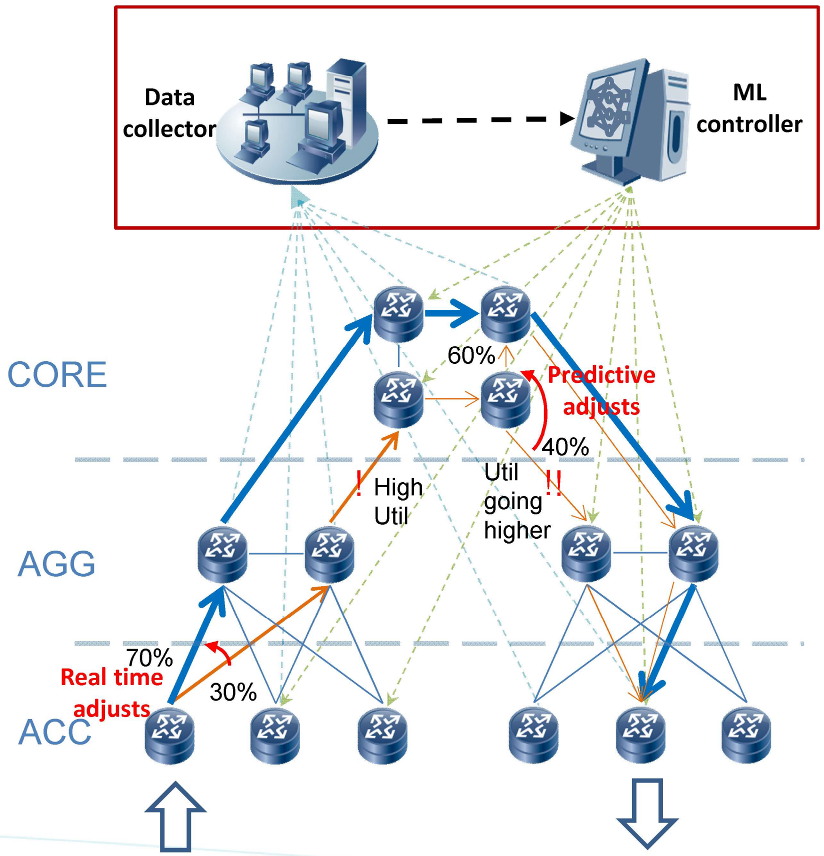
✓ **Congestion & Packet Dropping Minimized**



✓ **Latency Distribution: 10～100 ms Latency from 10% to 0.9%**

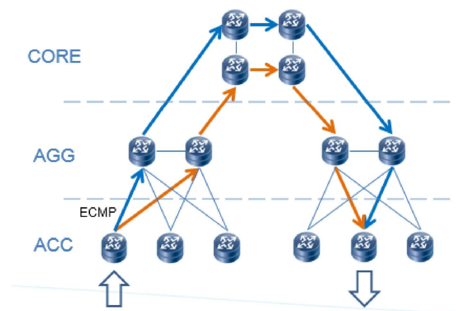


Case Study 2: Traffic Balance (Solution)



Simulation of dynamic traffic balance in MAN

VS



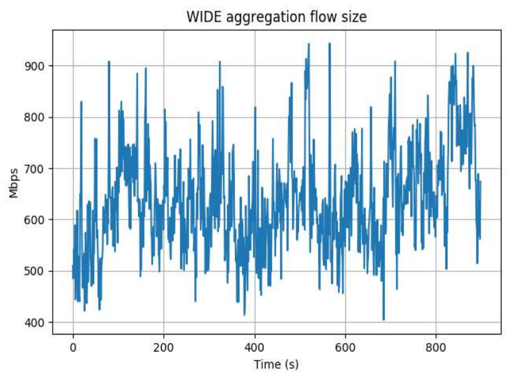
Simulation of ECMP in MAN

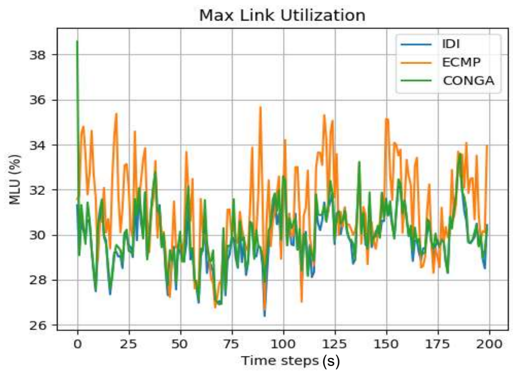
ECMP/UCMP drawbacks:
- No response to dynamic congestion
- Hashing by flow suffers from unbalanced traffic caused by elephant flows

Solution

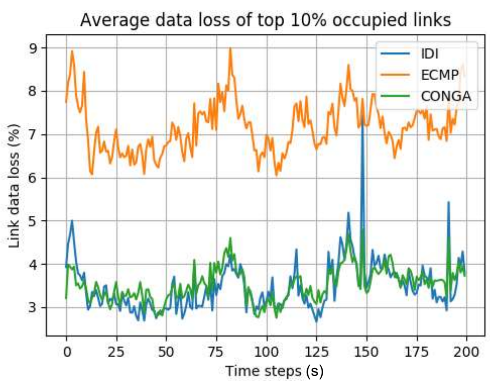
- Multi-path route computation by traditional Graph Theory algorithm, for each IE pair.
- Online Reinforcement Learning algorithm(DDPG) adjusts traffic distribution at every router to multiple outgoing paths.
  - Objective is to minimize Max-Link-Utilization.
  - Output is percentage of traffic over each path for each IE flow
- Modified forwarding plane forwards packet according to percentage, in flowlets to keep packet order

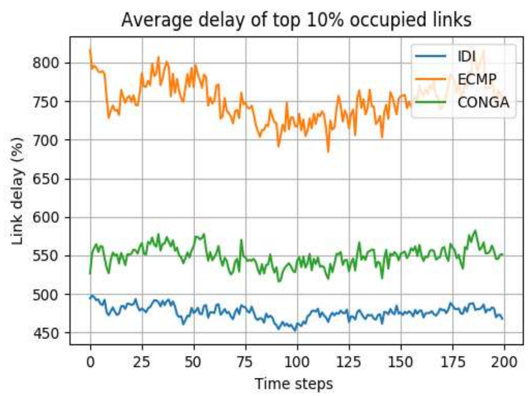
Case Study 2: Intelligent Traffic Balance (Demo Testing)

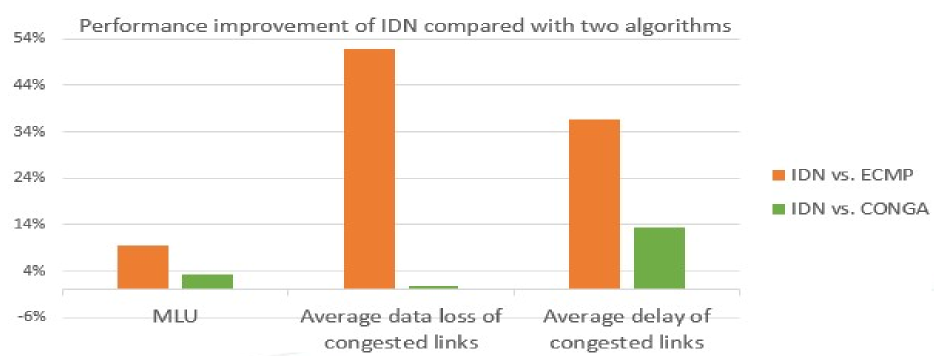




- Reducing Max-Link-Utilization by 23% (vs ECMP)
- Reducing data loss by 64% (vs ECMP)
- Reducing link latency by 19% (vs ECMP)
- Achieving close performance on both MLU and packet loss with CONGA(**ms-level** distributed control) by **second-level** centralized control, and better performance on delay
- Automatically adapts to evolving traffic pattern and network status



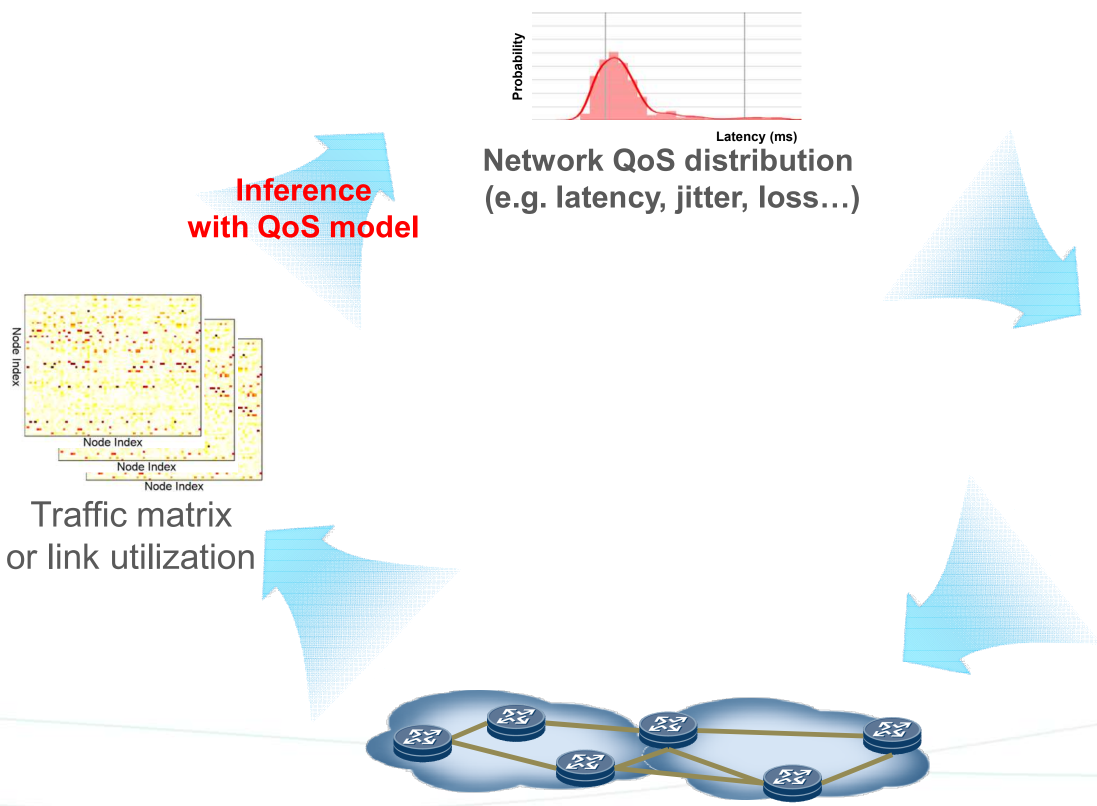




Case Study 3: QoS Modeling -- Insight & Prediction from Flashing Randomness

Problem:

If precise prediction of QoS values (**latency, jitter, loss...**) is impossible, can we predict the probability distribution of QoS? Conditioned by traffic rate



Application examples:

*1. Visualize & Analyze the **Historical/Real-time** QoS*

2. Predict ***Future*** *QoS*

Input estimated future traffic rate, predict the future QoS distribution of selected network paths.

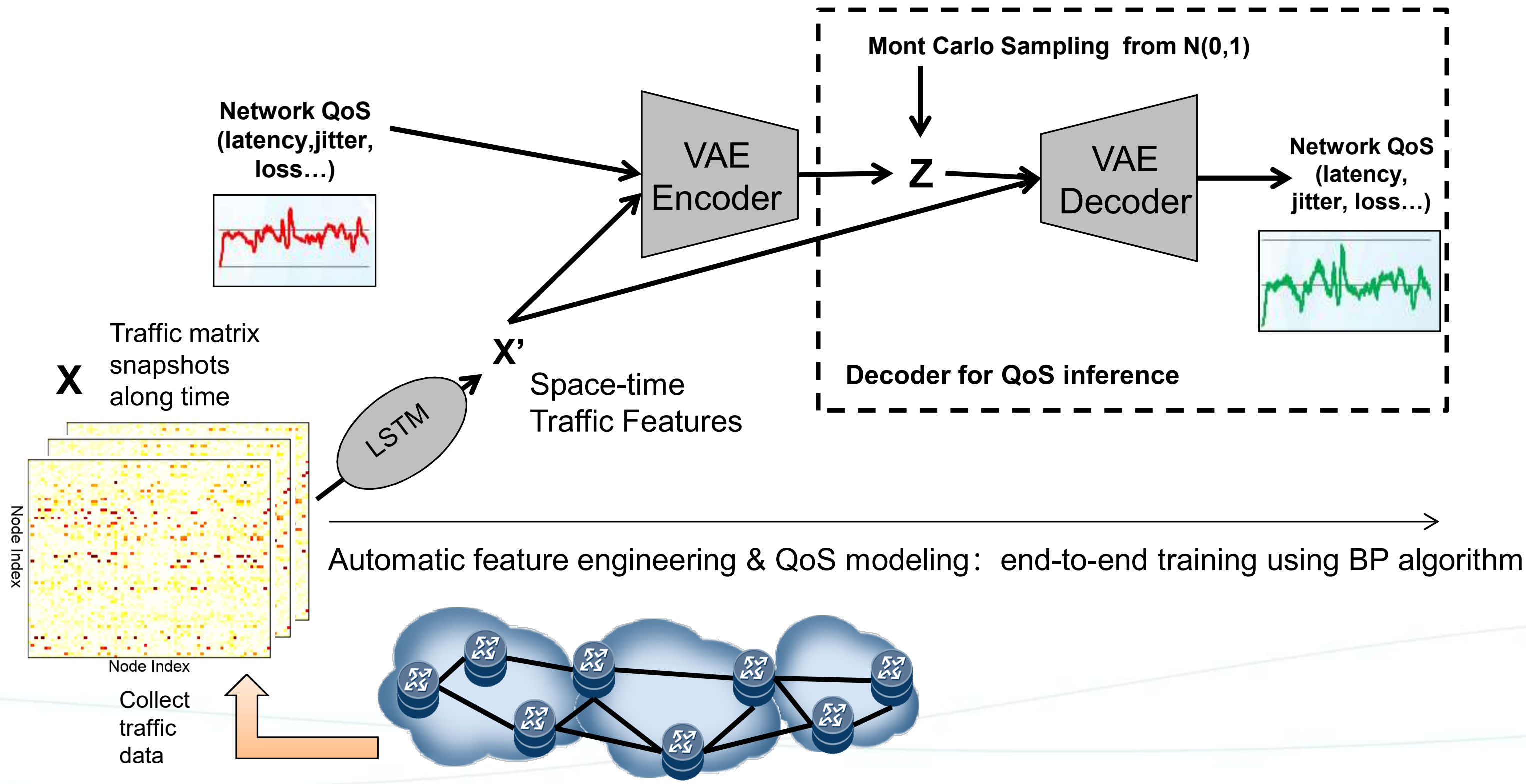
3. Decision-supporting in ***open loop control***

- Find path for QoS-critic service flow
- Schedule time-range routing policy for QoS-critic service flow
- Network capacity expansion planning under QoS restrictions

4. Additional input to, or output-evaluation in ***close loop control***

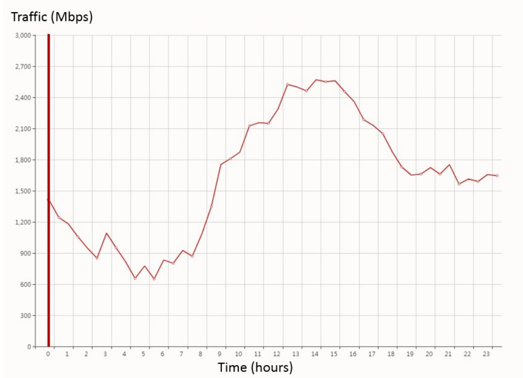
Case Study 3: QoS Modeling (Solution)

Method: By training with the historical traffic and QoS samples, the ***LSTM module*** learns to extract the traffic features and the ***VAE module*** learns the mapping from traffic features to QoS distribution. After training, it can ***predict the full distribution of various QoS metrics under any given traffic loads in real time with a high accuracy.***

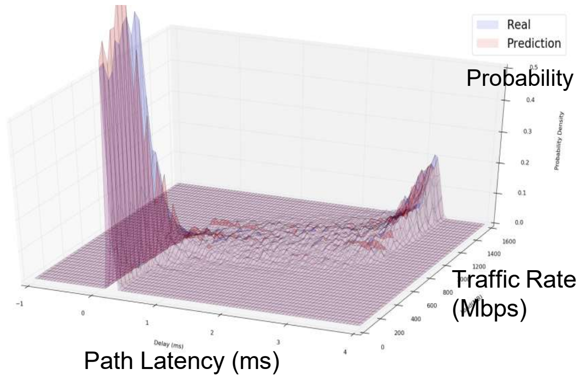


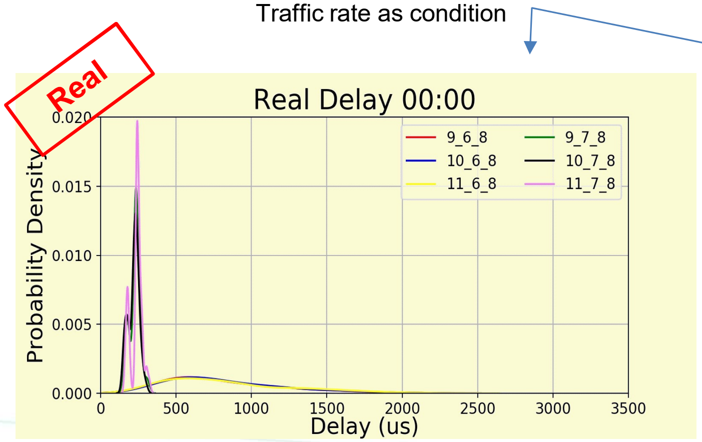
Case Study 3: QoS Modeling (Demo Testing)

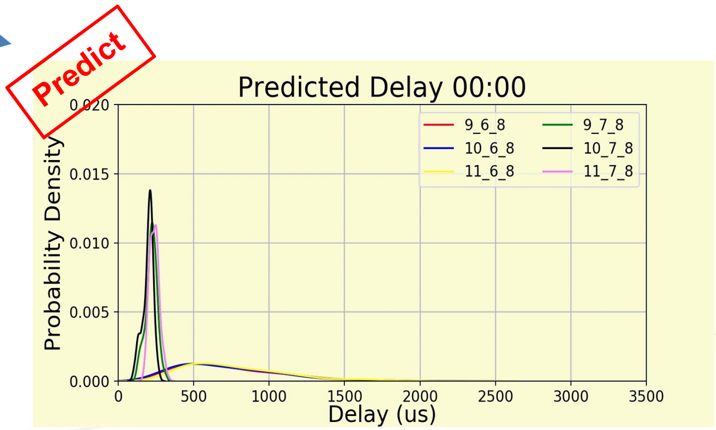
Results: 24-hour WIDE traffic data over test bed, parallel prediction of latency distribution of 6 network paths



Traffic rate as condition

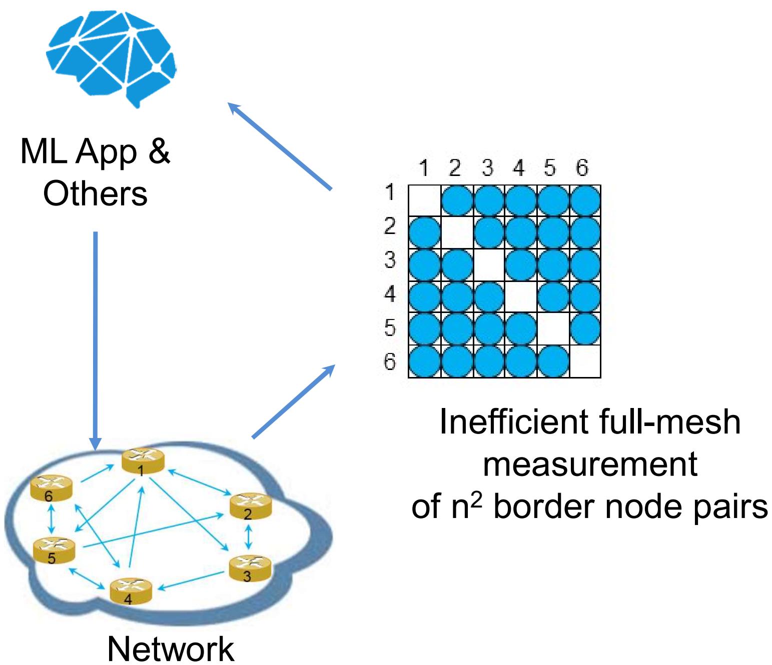






Predicted latency distribution of multiple paths

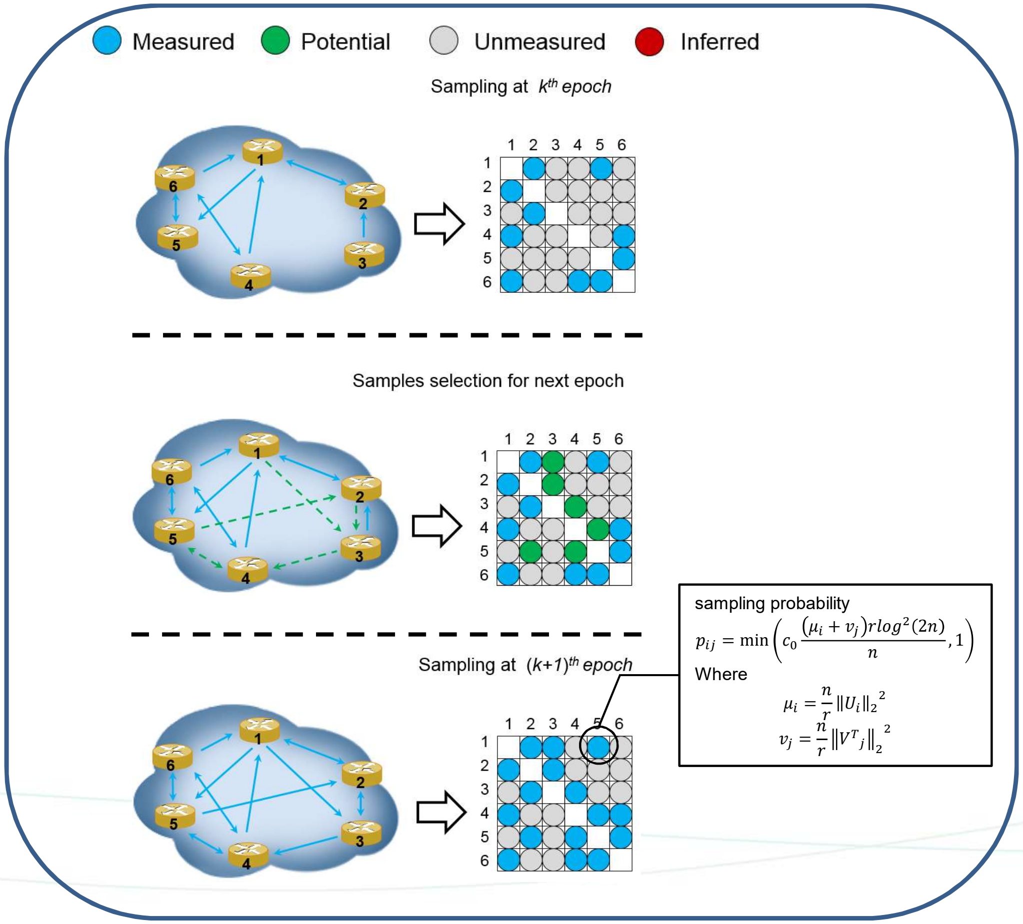
Case Study 4: Intelligent measurement (Open Issues)



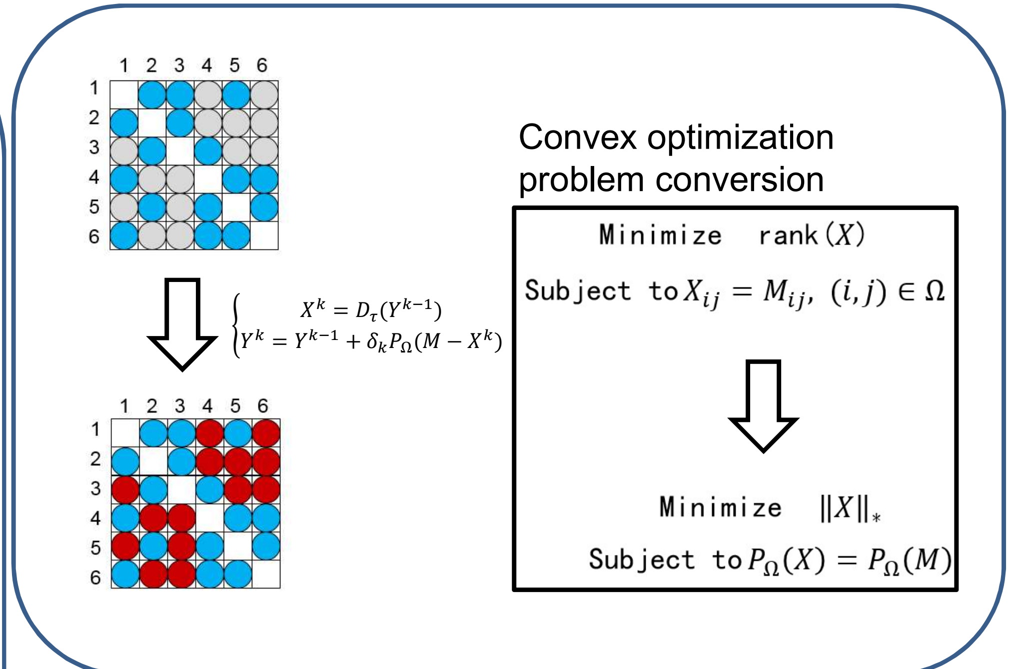
- "Data thirsty"
  - Fulfillment of tremendous-data requirement for intelligent applications and network insight
  - Network visualization to gain the vision of network performance
  - Capability of measurement for large-scale networks (TB data per day)
- Problem
  - Infeasible for full-mesh measurement between all transmission pairs in practice
  - Sampling small portion of transmission pairs to recover information via matrix completion
- Challenges
  - Traditional sampling is inflexible with fixed rate, which cannot be done on-line
  - Exact matrix completion requires uniformly random sampling and incoherent assumption
  - Theoretical lower limit for exact matrix completion can hardly be found in practice

Case Study 4: Intelligent measurement (Solution)

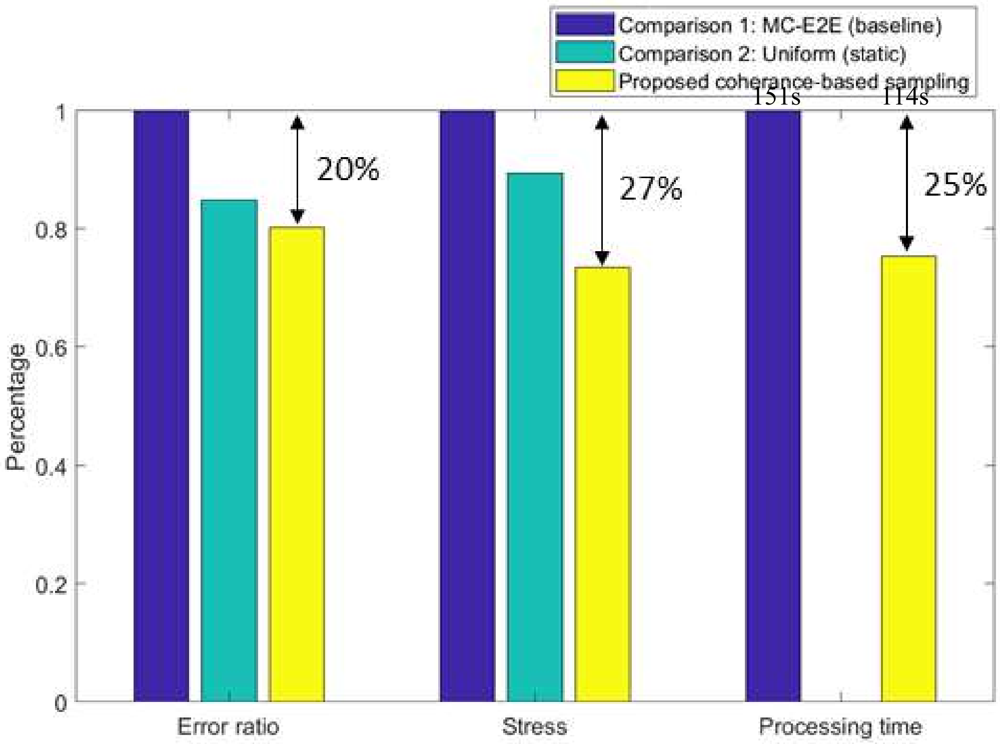
1 Coherence-based dynamic sampling



2 SVT-based matrix recovery



Case Study 4: Intelligent measurement (Demo Testing)



Performance improvement on latency measurement

- Sampling rate: 20%
- Accuracy: 94.8%
- Error ratio:

$$\frac{\sum_{i,j=1}^{n}\left|X_{i,j}-\hat{X}_{i,j}\right|}{\sum_{i,j=1}^{n}X_{i,j}}$$

- Stress:

$$\sqrt{\frac{\sum_{i,j=1}^{n}(X_{i,j}-\hat{X}_{i,j})^2}{\sum_{i,j=1}^{n}{X_{i,j}}^2}}$$

- Processing time: Note that Comparison 2 of uniform sampling is static without online scheduling

- Data set：Harvard226. Latency among 226 nodes

Content

What can ML bring to IP Networks?

Applications

ML Engineering

Intelligent Network Infrastructure

ML Engineering: from Academy to Industry

2030s — “AI is eating the world”?

2020s — AI/ML Engineering?

2010s — AI/ML as “ black magic”

2000s — “Software is eating the world”

1970s — Software Engineering

1960s — Software Crisis

1950s — Software as “ black magic”

ML Engineering for network industry

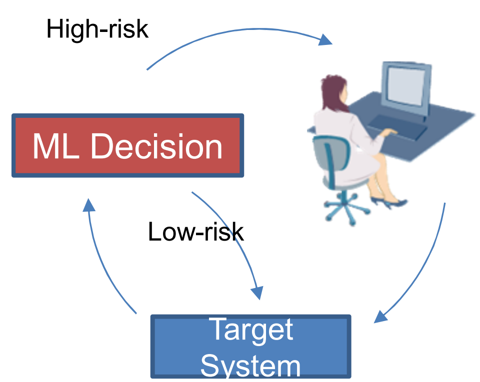
- Robust ML: sand box, threshold limitation, robust online-learning …
- White-box ML: Interpretability
- Fast-starting: smoothen the impact of the learning phase.
    - few-shot learning
    - transfer learning
    - hybrid control
- Scalability: converge over large environment
- ML in forwarding plane
    - computation @ wire speed
    - compact model in forwarding chips, e.g. integer, even binary parameters in ANN

ML Engineering: Robust ML

ML as a nondeterministic system, may produce unacceptable decisions with certain probability

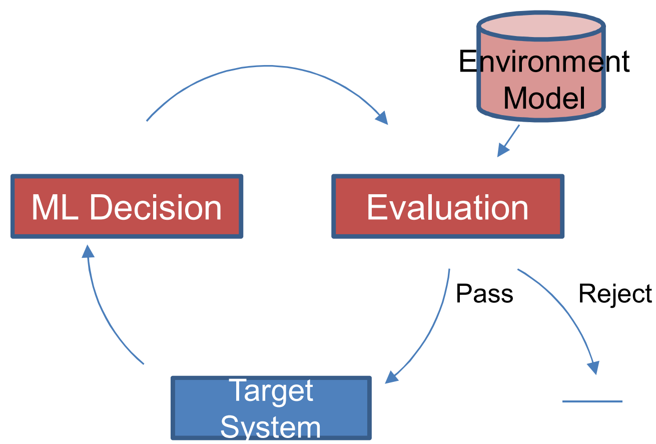
Even though it is unlikely be fully resolved in near future, there are engineering designing to allow ML application in critical scenarios

Human-in-the-loop

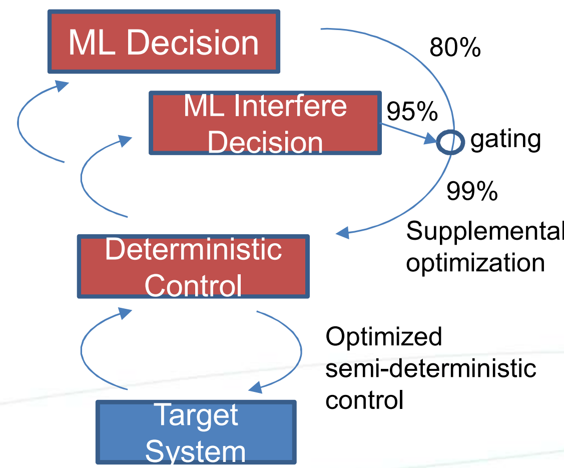


Example: Amazon customer service system

Close loop with evaluation

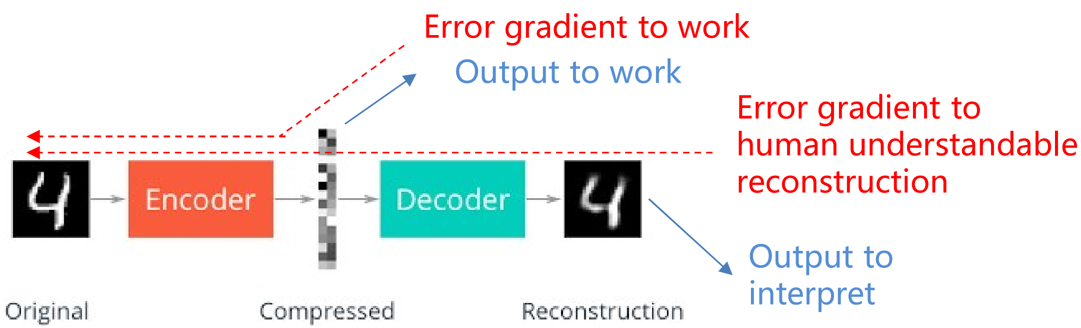


Supplemental control loop

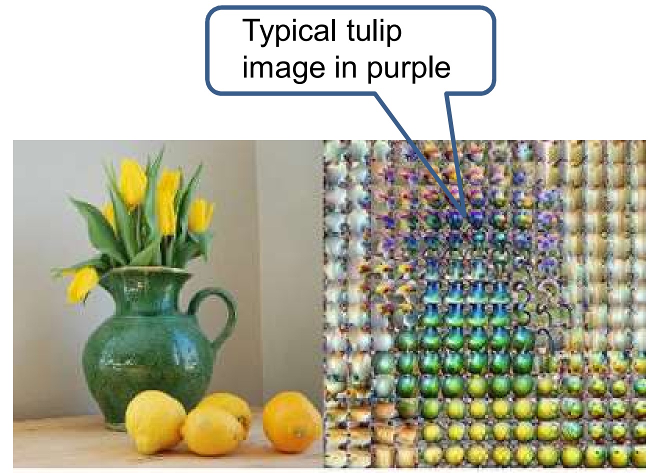


ML Engineering: White-box ML

The popular belief that AI/ML is a black-box, is challenged by recent research



Dual training flows: additional Auto-Encoder to train interpretable ANN



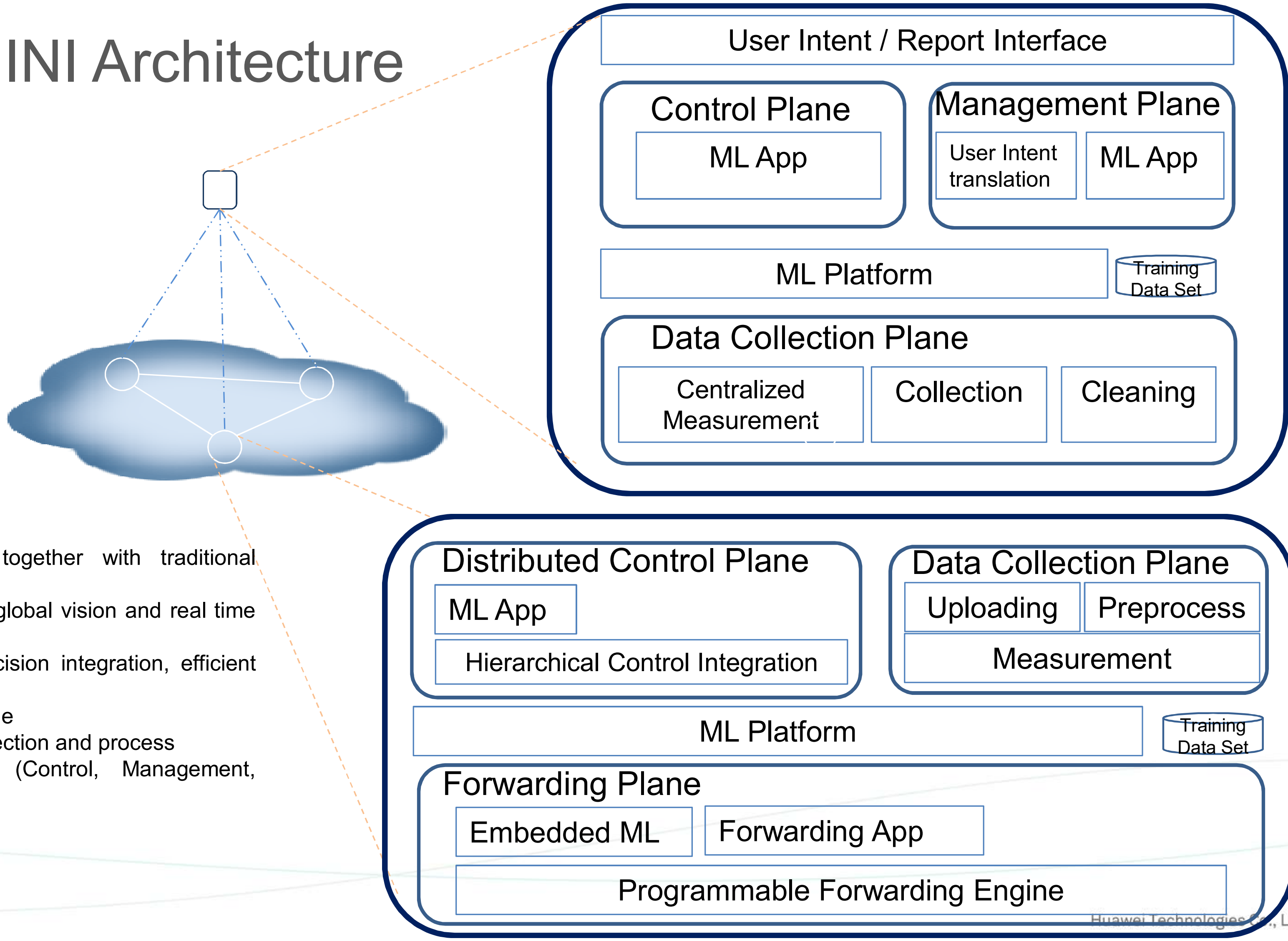
On the left is an image that was put through a neural network trained to classify objects in images — for example, to tell whether an image includes a vase or a lemon. On the right is a visualization of what one layer in the middle of the network detected at each position of the image. The neural network seems to be detecting vase-like patterns and lemon-like objects.

Google: The Building Blocks of Interpretability

Find out the meaning of a neuron by examining:
Which images activate this neuron?
Which patches in the image activate this neuron?
Which classes does this neuron helps to classify?
Choose a typical image from the major class to mark this neuron.

Content

Next Step: INI Architecture



- ML in every plane, working together with traditional functionality and protocols
- AI in every entity, covering both global vision and real time actions
  - Cooperation issues: decision integration, efficient communication
- A new plane: Data Collection Plane
  - Data measurement, collection and process
  - Serving three planes (Control, Management, Forwarding)

Q&A

Thank You.