Diagnosis in Practice

Rick Schlichting

Executive Director, Software Systems Research
AT&T Labs
Florham Park, NJ 07932, US

Special thanks to Matti Hiltunen and Jen Yates from AT&T for slides and advice on the talk.

Diagnosis

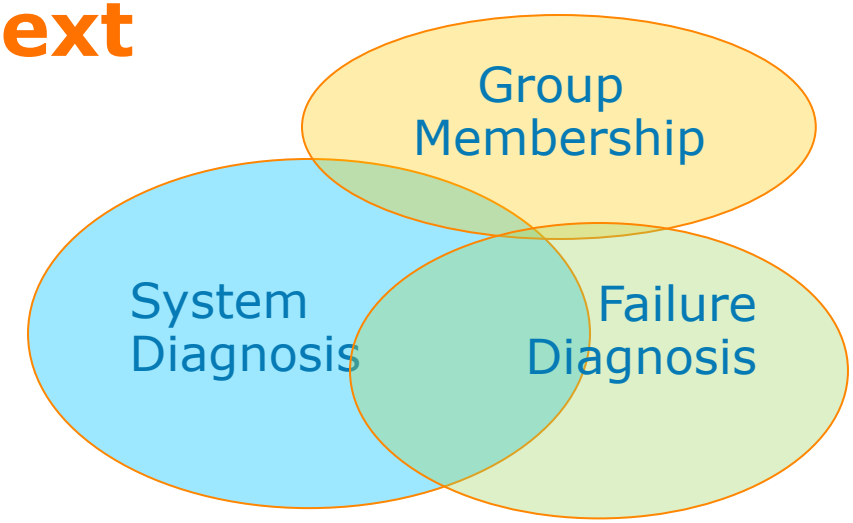
Three fundamental steps:

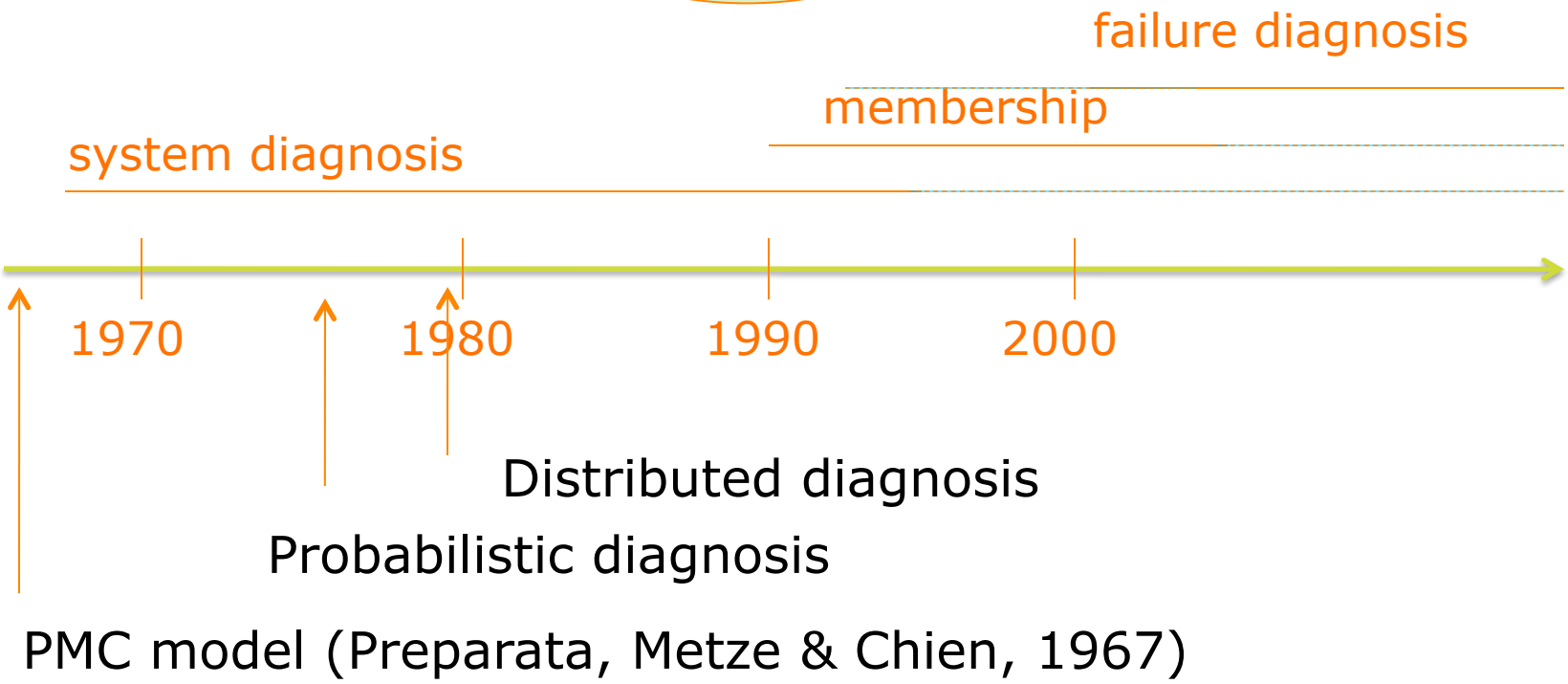
1. Detecting that something is wrong
2. Figuring out what is causing the problem (diagnosis, finger pointing, root cause analysis)
3. Fixing it (repair)

But — *lots* of questions and issues in realistic networked systems:

- What are "something" and "wrong"?
- Is the source of the problem under my control?
- Is it possible for me to fix the problem?
- How to deal with scale?
- ….

Context





Talk

Focus on enterprise- and ISP-scale infrastructure and services

Two themes:

- Reality
- Two example research projects that attempt to deal with the challenges

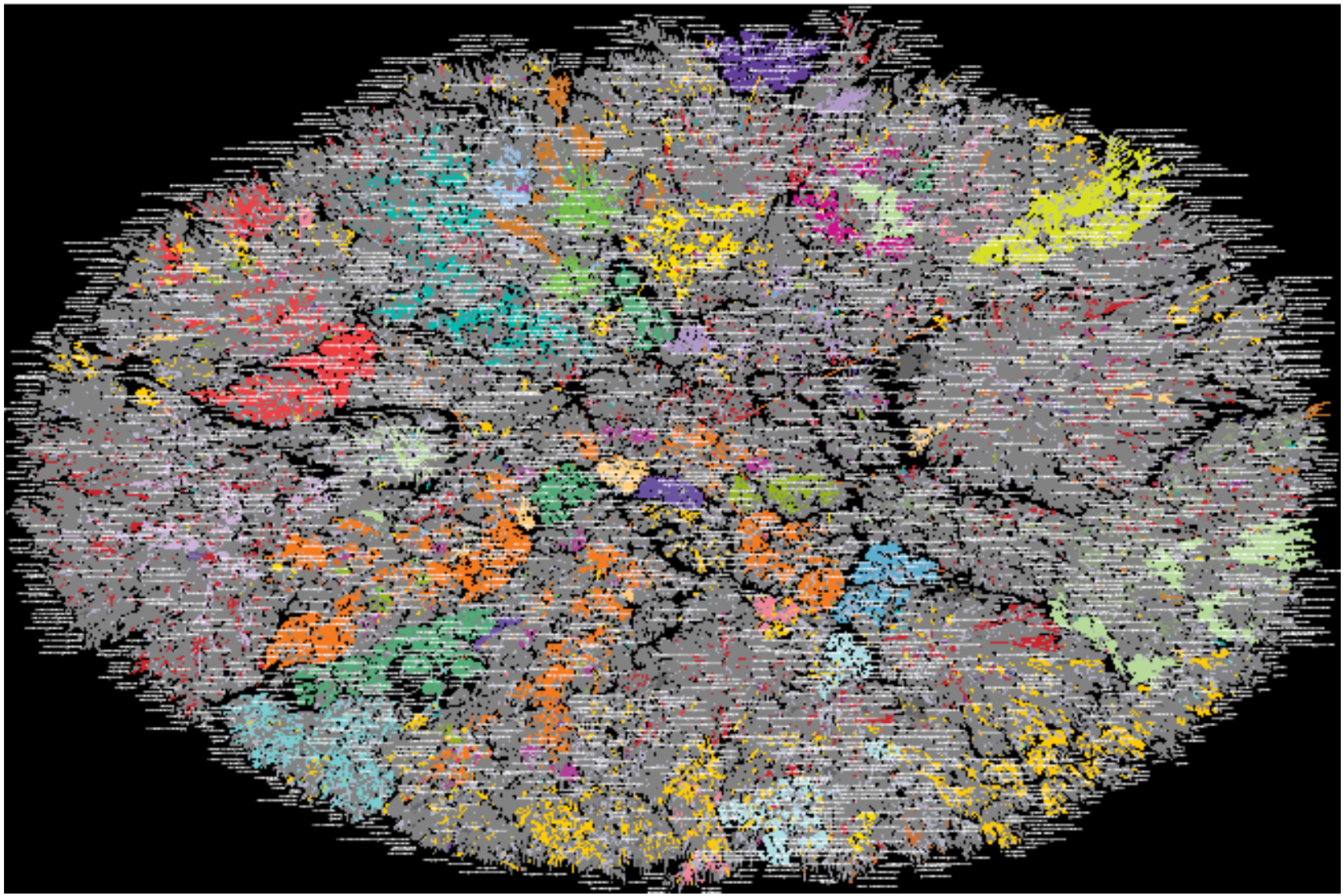
Goals:

- A better understanding of the issues involved
- Energized to work on the challenges!

So What's the Reality?

- Scale and Complexity

The Visible Internet

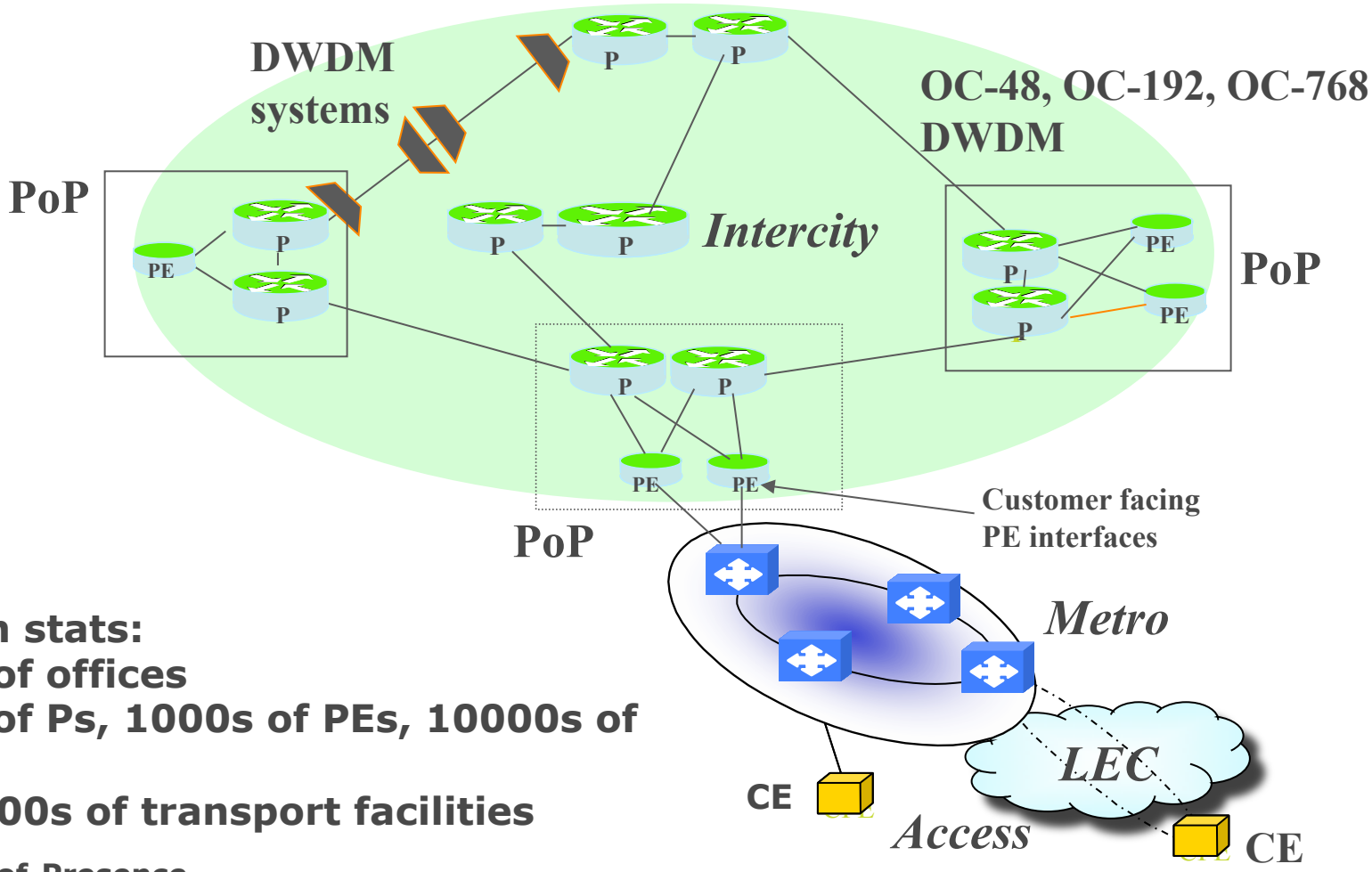


Global Networks (AT&T)

Over 20 Petabytes of Traffic Average Business Day

MPLS-based Services in 140+ countries at 3800+ service nodes
35+ data centers on 4 continents
Over 875,000 route miles

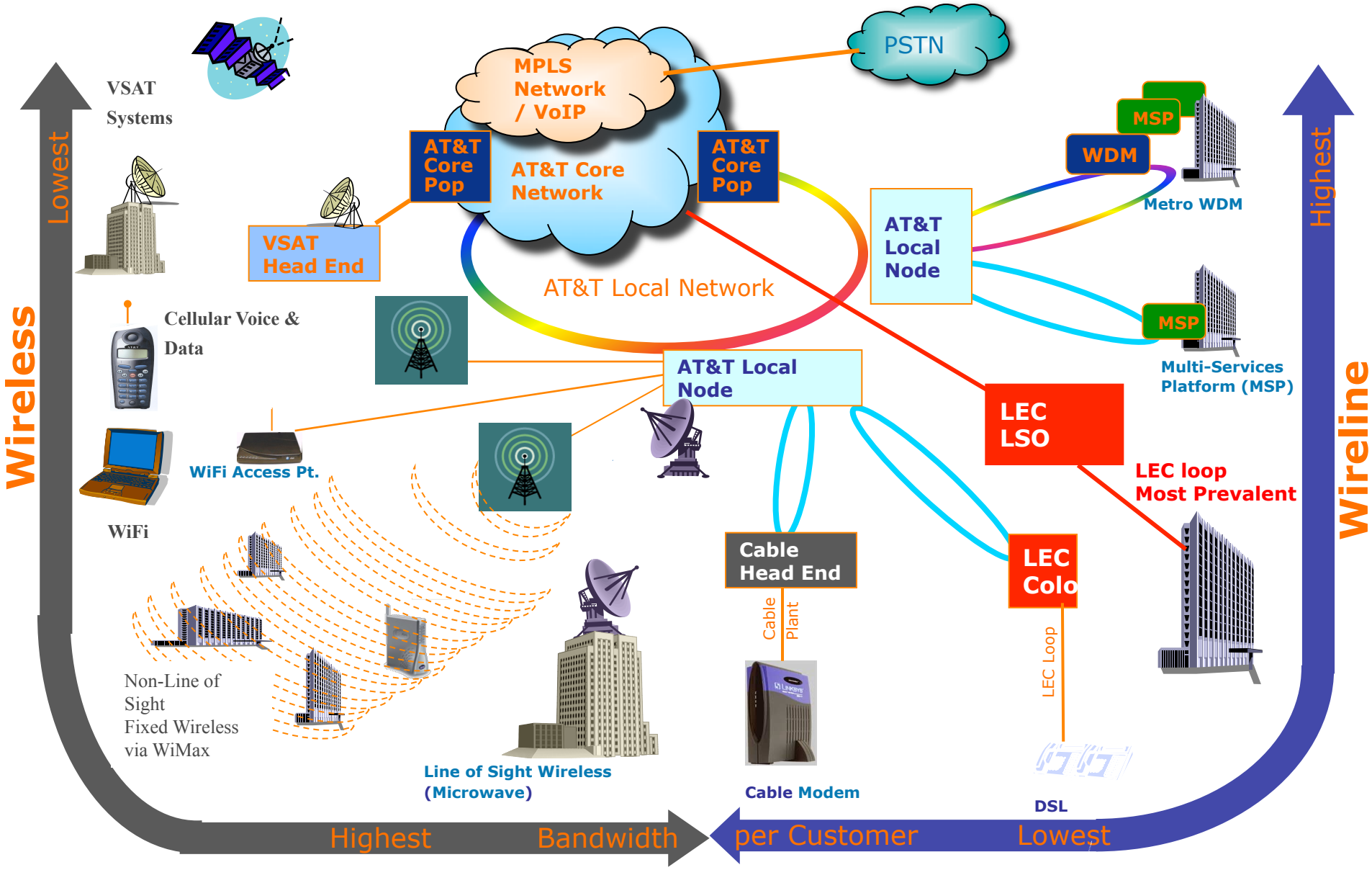
IP Networks



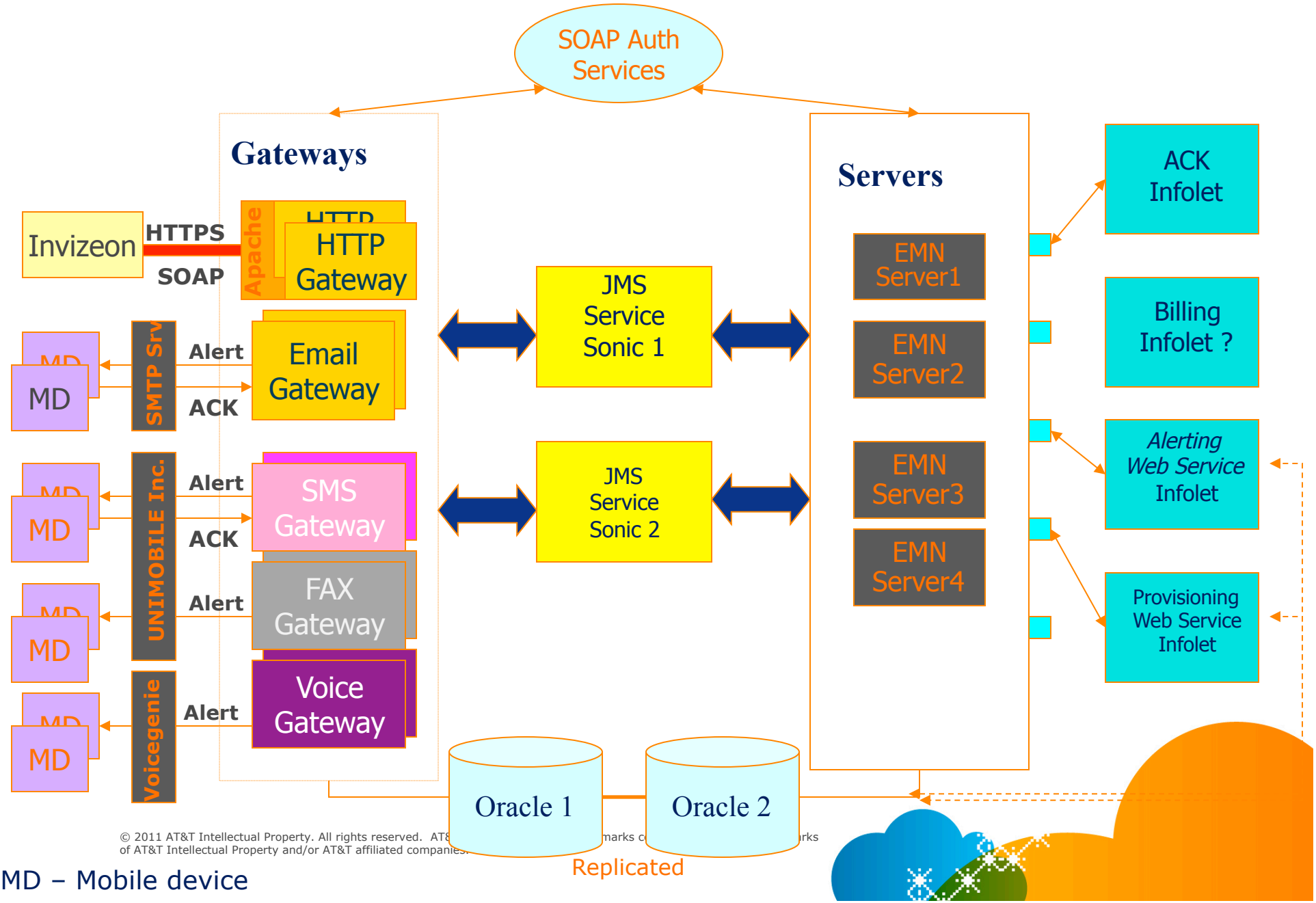
Rough stats:
100s of offices
100s of Ps, 1000s of PEs, 10000s of CEs
100,000s of transport facilities

PoP: Point-of-Presence
P: Backbone (core) Router
PE: Provider Edge Router
CE: Customer Edge Router

Access Networks



Enterprise Applications



MD – Mobile device

Execution Infrastructure

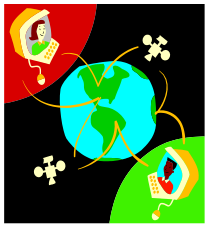
Dynamic shared environments

- Cloud computing and utility computing.
- VMs

No longer as simple as server + OS

- Layers upon layers of software

Increasing Complexity

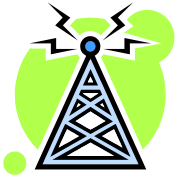

Massive scale
100s of offices, 1000s of routers,
10,000s of interfaces,
Millions of consumers

Immense software complexity
Scale, Bugs, Interactions

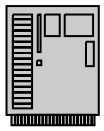




Applications
Scale, sensitivity


Diverse technologies and vendors
Layer-1, Layer-2, Switches, Routers,
IP, Multicast, MPLS, wireless access points




Continuous evolution
Upgrades, Installations

Other Practical Implications

Size:

- Potentially multiple independent problems may be present in the system at the same time
- Data volume

Data quality:

- Limited forms of event data available (e.g., trouble tickets, monitor outputs, some logs, only failure data)
- Data available for only part of the system
- Gaps in data, delay in data collection
- Correlation of data from heterogeneous systems

So What's the Reality?

- Spectrum of issues

AT&T Network Disaster Recovery

Specially Designed **Recovery** Equipment

More than a $0.5 Billion investment in over 600 trailers and support vehicles containing network technology, infrastructure, and support elements.

Security Network Traffic Analysis

Customer Business Issues

Issues and Challenges:

- Networks are becoming more complex
  - Thousands of vulnerable targets and hundreds of applications
  - Different threats such as internal misuse, zero-day attacks, etc.
  - Variety of security services 24x7 at low cost
- Disruptive network attacks could negatively impact business

Solution should address:

- Ability to analyze security network traffic
- Alert and Notification functionality
- Capacity to track and report historical traffic
- Minimum "Total Cost of Ownership" solutions

The AT&T Internet Protect® Service Family

	AT&T Internet Protect®	My Internet Protect	DDoS Defense	Private Intranet Protect
Scope	• Security alerts & notification for identified threats impacting AT&T IP Network	• Security alerts & notification for identified threats entering customer's network	• Security alerts and mitigation of identified DDoS attacks impacting customer's network	• Security alerts, notification, and analysis for identified threats within customer's network
Features	• Notification via email/ paging • Mitigation recommendations	• Tracks private network IP addresses • Helps detect profile and misuse anomalies	• Tracks private network IP addresses • Helps detect profile and misuse anomalies • DDoS mitigation	• Alerts based on unsampled network data
Benefits	• Early warning of identified possible attacks • Requires no additional hardware or software	• Measure anomaly traffic volume flowing towards the private network • Requires no additional hardware or software	• Designed to help pro-actively eliminate DDoS attacks before they penetrate private network • Requires no additional hardware or software	• Perform security analysis within and outside private network • Requires no additional hardware or software
Reports	• Alert Summary • Alert Details • Port Summary	• Traffic Summary • TCP Application Summary • UDP Application Summary	• Traffic Summary • TCP Application Summary • UDP Application Summary	• Host Details • Service Details • Group Details • Compliance with customer's security policies
Requirements	• None	• AT&T Internet Protect® • MIS, or GMIS	• AT&T Internet Protect® • MIS	• AT&T Internet Protect® • EVPN

Failures at different layers

- Network, hardware, software, software configuration

Different failure types

- Crash, performance, omission, quality, value, …

Permanent and transient failures

Chronic failures

- low impact, repeating

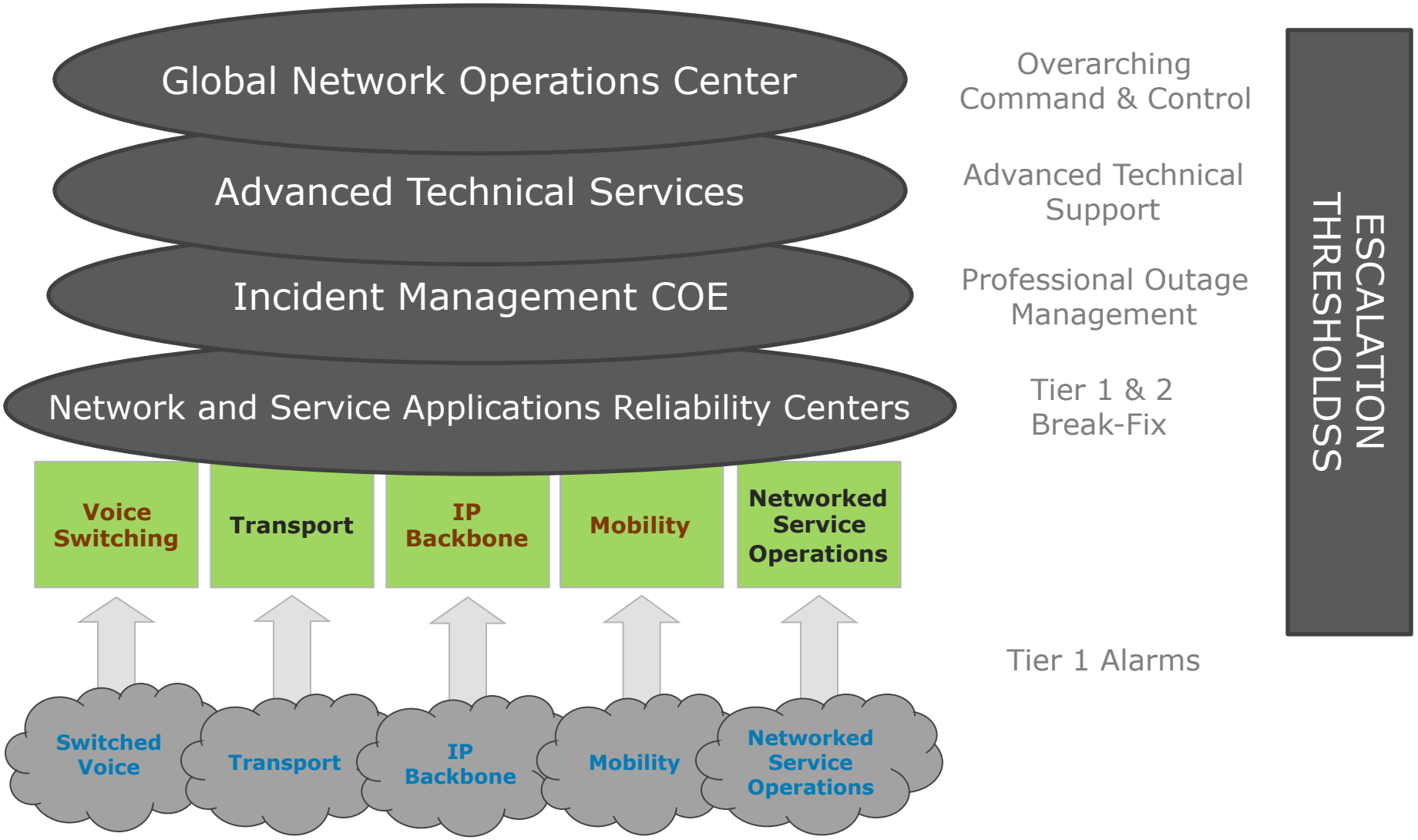
➥ Focus increasingly on *service quality management* (SQM)

- individual component or network element -> end-to-end service quality
  - Machine, router, link, line-card -> User perceived performance in a CDN or VPN service
- hard failure -> transient problem
  - Fiber cut, router failure, crash -> transient performance degradation, protocol flap

So What's the Reality?

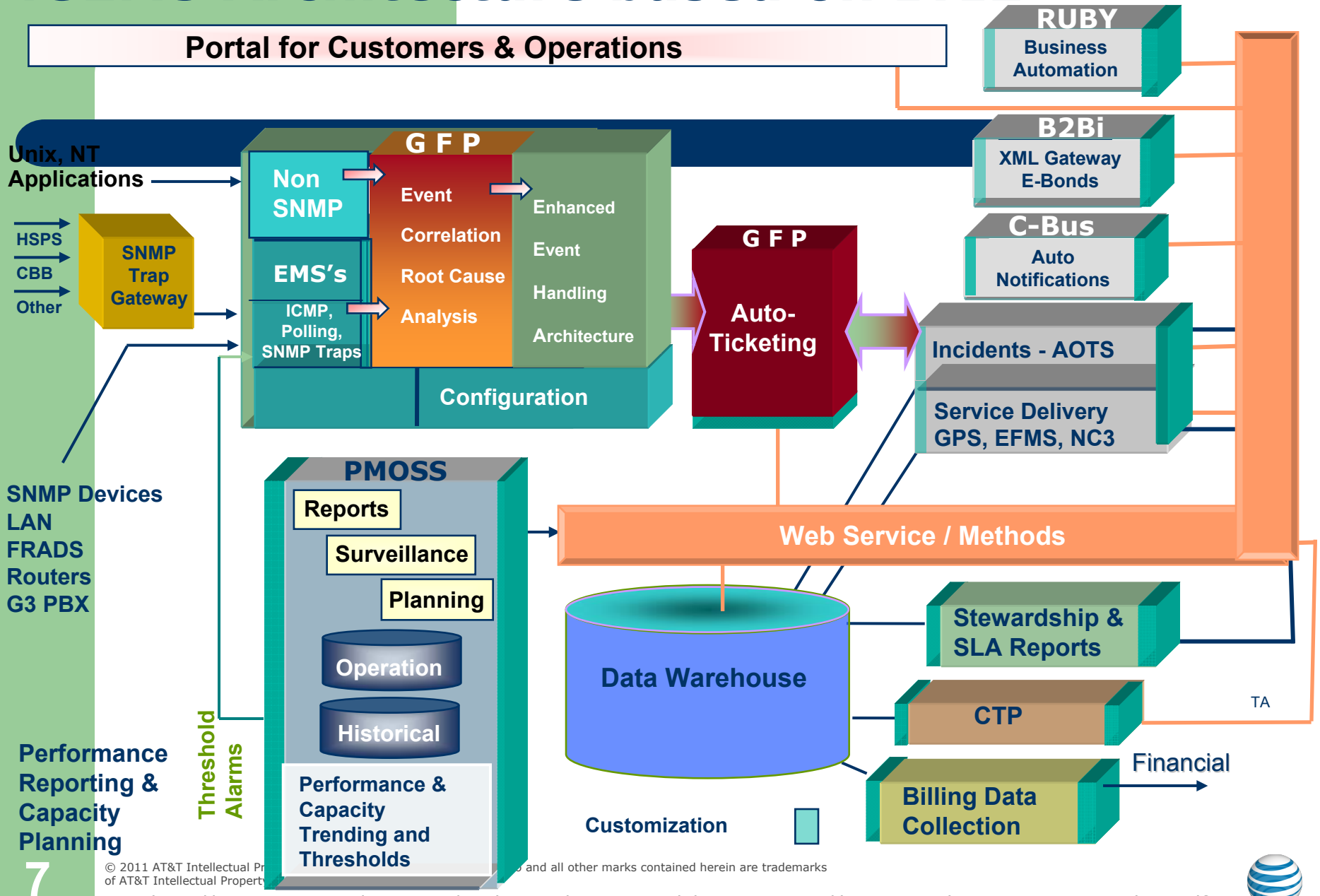
➥ **Current practice**

Global Network | Center-based Operations



AT&T Global Networks Operations Center (GNOC)

iGEMS Architecture based on ITIL



http://www.naspi.org/meetings/workgroup/2008_march/presentations/ibm_network_managment_nicoletti.pdf

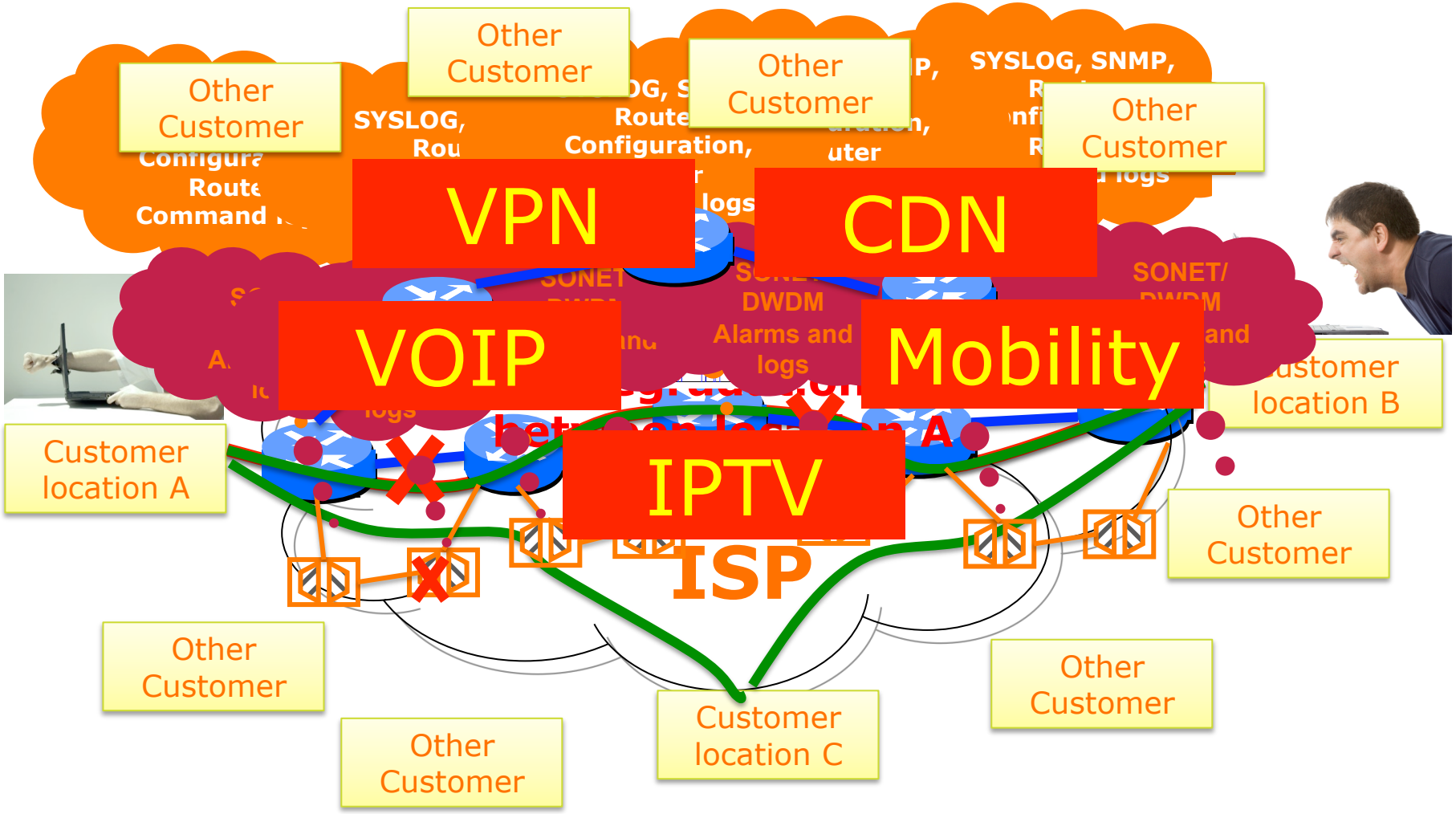
G-RCA: A Generic Root Cause Analysis Platform

He Yan[1], Lee Breslau[2], Zihui Ge[2], Dan Massey[1], Dan Pei[2], Jennifer Yates[2]

[1]Colorado State Univ, [2]AT&T

ACM CoNEXT 2010, Nov 30-Dec 3, 2010, Philadelphia, PA, US

Service Quality Management (SQM) and Root Cause Analysis (RCA)



Challenges in RCA for SQM

Complex service dependency model

- The quality of a VPN connection across the ISP network depends on the status of the routers and links along the network path carrying the traffic, which is dynamically determined based on the link weights

Ever-changing environment

- New services (e.g., multicast VPN), new technologies (e.g., MPLS TE), new devices (e.g., OC768 line cards) are introduced into ISP networks at a fast rate
- Operators need new RCA tools for new services
- RCA tool needs to deal with imperfect domain Knowledge

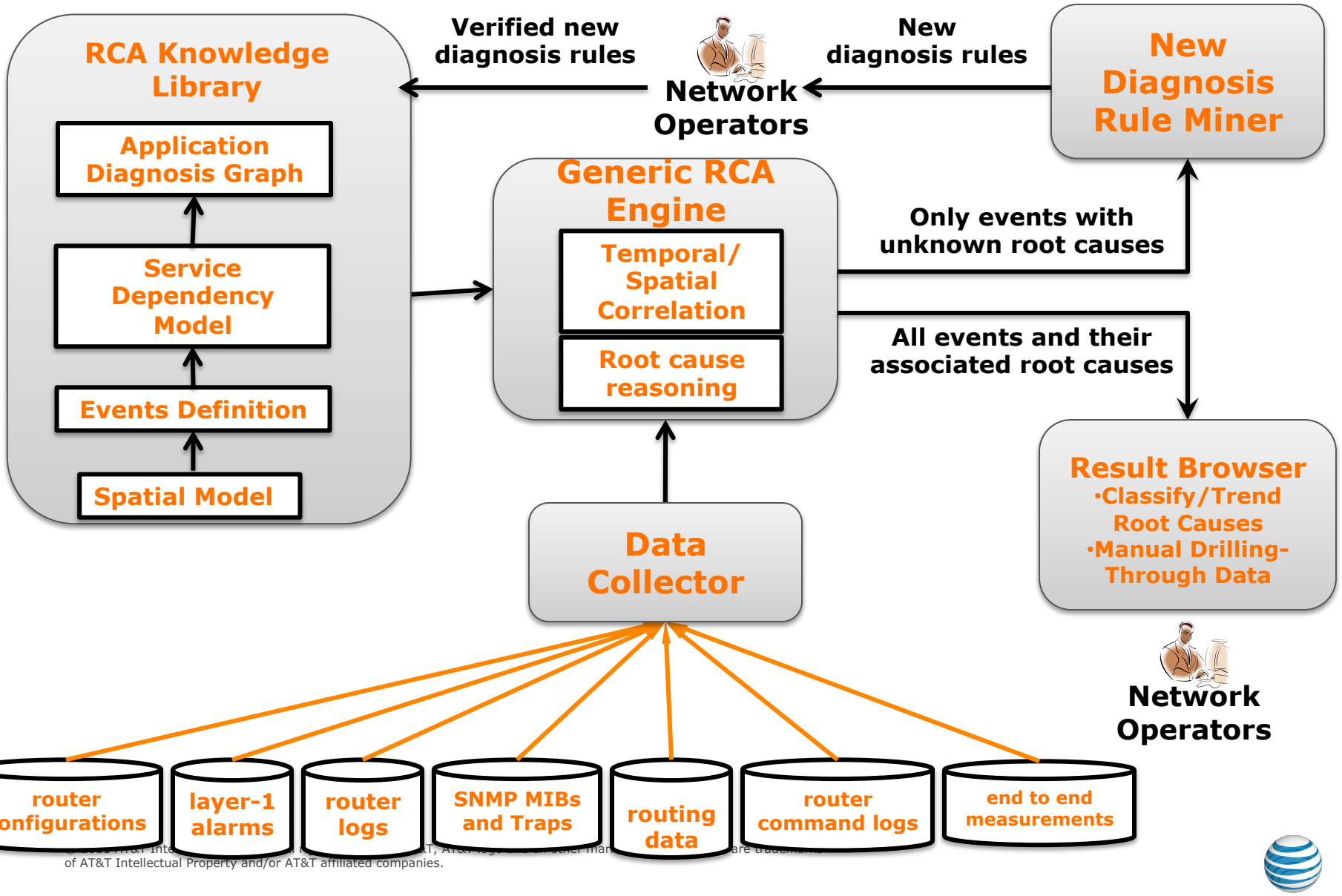
Data Collection

- Proactively collect data (alarms, logs and performance measurements) to enable the analysis of historical transient service disruptions

Scalability

- Analyze a large number of transient service disruptions over extended period to determine the primary root causes
- Dig into millions of alarms and logs distributed all over the network.

G-RCA Design



RCA Knowledge Library

➥End result is Application Diagnostic Graph

Spatial Model

- Topological information, cross-layer dependency, logical and physical device association
- Dynamic routing
- Mapping is time dependent

Event

- A signature that captures a particular type of network condition, e.g. Interface flap or router reboot
- G-RCA pre-defines and implements a wide range of commonly used events.

Service Dependency Model

- Consists of *diagnosis rules*
- Symptom and diagnostic events are picked from event definitions, e.g.: interface flap -> line protocol flap
- G-RCA pre-defines the commonly used diagnosis rules

Generic RCA Engine

Temporal Correlation Module

- Given a symptom event instance, find out the temporally correlated the diagnostic event instances
- Imperfect timing

Spatial Correlation Module

- Given a symptom event instance, find out the spatially correlated diagnostic event instances
- Indirect mapping

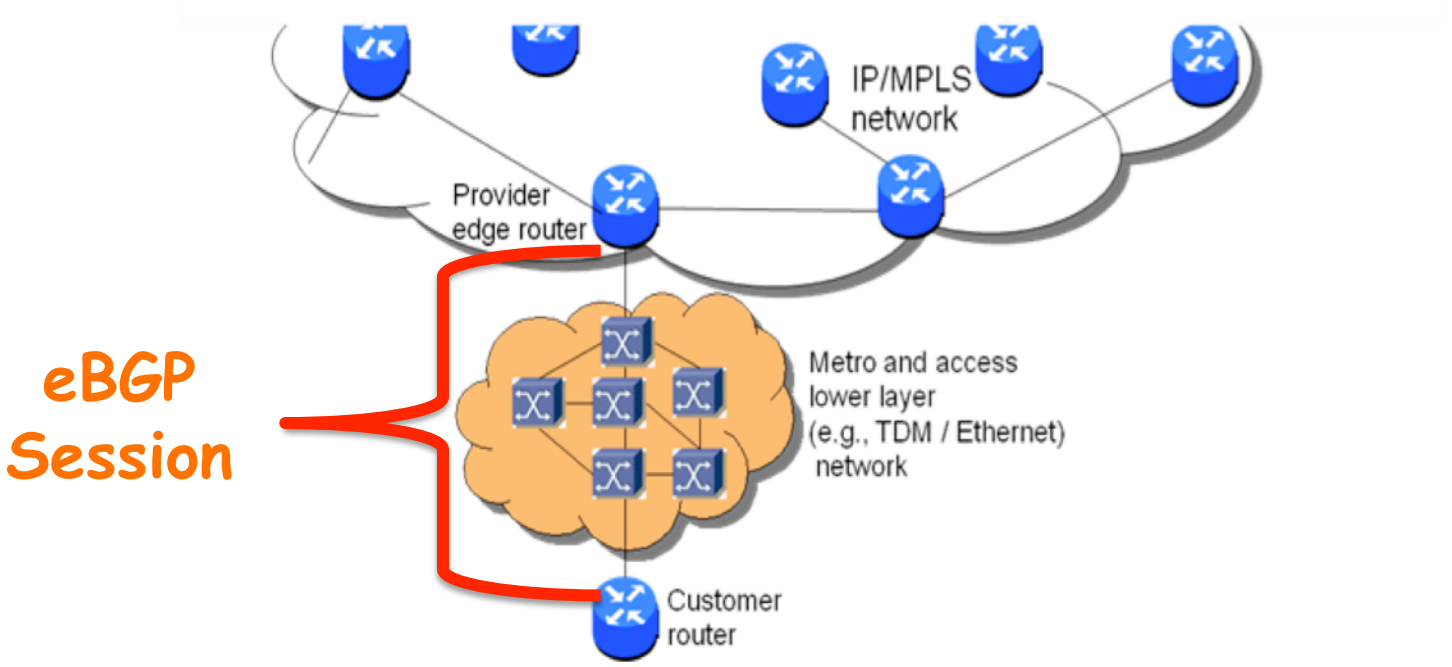
Root Cause Reasoning Module

- Determine the root cause of a symptom event instance based on temporally and spatially correlated diagnostic event instances.
- Rule-based decision-tree-like reasoning
  - associate a priority value for each edge in the diagnosis graph
  - The higher the priority value, the more likely the diagnostic event to be the real root cause

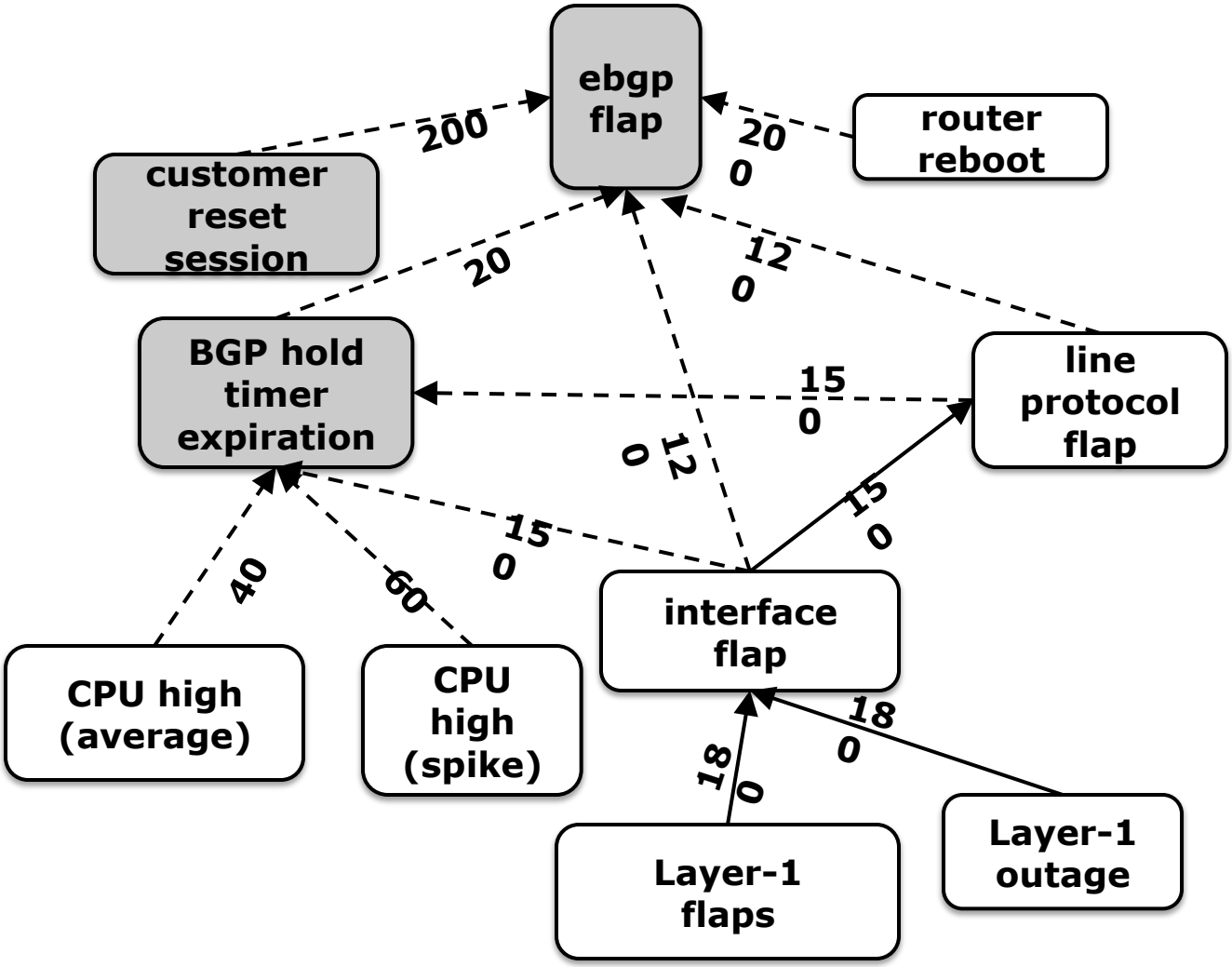
Example: BGP flaps root cause analysis

Minimize the number of eBGP session flaps

A challenging problem across the trust domain



Application Diagnosis Graph



Operational Experience

5 provider edge routers, each with several hundred eBGP sessions with customer routers

An interesting discovery!

Root Cause	Percentage (%)
router reboot	0.047
customer reset session	0.088
CPU high	0.886
CPU rising (high)	15.32
CPU rising (medium)	4.318
interface flap	29.004
line protocol flap	15.72
eBGP HTE(due to unknow reasons)	18.381
Short Layer-1 Flaps	0.205
Longer Layer-1 Outage	0.428
unknown	15.603

Key Take Away

A generic, automated and scalable root cause analysis platform for SQM in large IP networks

Draco: Statistical Diagnosis in VOIP Systems

Soila Pertet Kavula[1], Kaustubh Joshi[2], Matti Hiltunen[2], Scott Daniels[2], Rajeev Gandhi[1], Priya Narasimhan[1]

[1]Carnegie-Mellon Univ, [2]AT&T

http://www.pdl.cmu.edu/Publications/

Motivation

Business VoIP services

- 10+ service types, 200+ network elements
- Millions of calls per day, growing rapidly

Faults occur continuously in the system

- Minor incidents + major incidents, e.g., failed upgrades, can increase fault rate
- Faults cause failed calls (blocked or cut-off calls)

Research question

- How to diagnose faults that have never happened before?
- Faults due to combinations of unexpected events?

VoIP SIP Call Flow

Call Setup

1. Calling party sends SIP invite

2. Get customer charge number & IP

3. Executes service logic.

IP Base Element (IPBE)

Call Control Element (CCE)

Application Server (AS)

4. Send routing number for destination

Gateway Server Exchange (GSX)

5. Send invite to destination. Call rings.

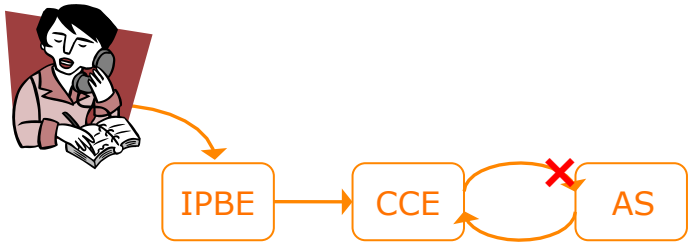
Data Transfer

6. Call established.

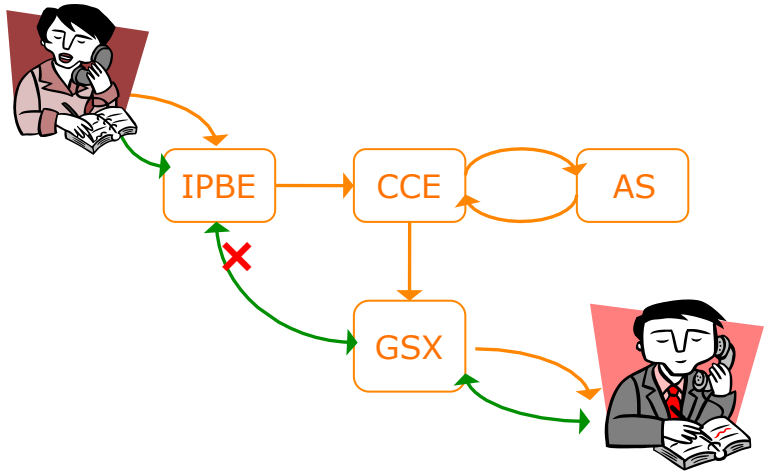
VoIP Call Detail Records (CDRs)

Network elements record call outcome in CDRs

- Timestamp, name of network element, service type
- Telephone numbers, customer IP addresses
- Outcome of call (successful, blocked, or cut-off call)



Blocked calls: Fail during setup



Cut-off calls: Failure after setup

Iterative Bayesian Approach

Adopt a technique from software debugging literature:

- Liu, C., Lian, Z., and Han, J. "How Bayesians Debug". In *Proc. 6th IEEE Int. Conf. on Data Mining. D*ec. 2006.

Operates on CDR "attributes"

- Service types, defect codes, network element names
- Customer IP addresses
- Easily extended to include additional call attributes
  - e.g. software versions, QOS data

➥Identifies call attributes most correlated with failed calls

Steps:

1. Extract call attributes from CDRs and model using a truth table
2. Compute distribution of each attribute in failed and successful calls
3. Iteratively select attributes that best differentiate failed from successful calls
4. Rank problems based on severity

“Truth table” Call Representation

Successful Call
- **ny4ny01sdh**
- **ph4pa0102bap**
- **ny4ny02gh**
- **….**

Failed Call
- **ny4ny01sdh**
- **at4ga03wap**
- **ny4ny02gh**
- **….**

ny4ny01sdh	ph4pa0102bap	at4ga03wap	ny4ny02gh	Call Outcome
1	1	0	1	SUCCESS
1	0	1	1	FAIL

Suspect Attribute Identification

In each call:

- Assume each attribute has a stable but unknown occurrence probability.
- Reflects service volume, call distribution, routing, etc.

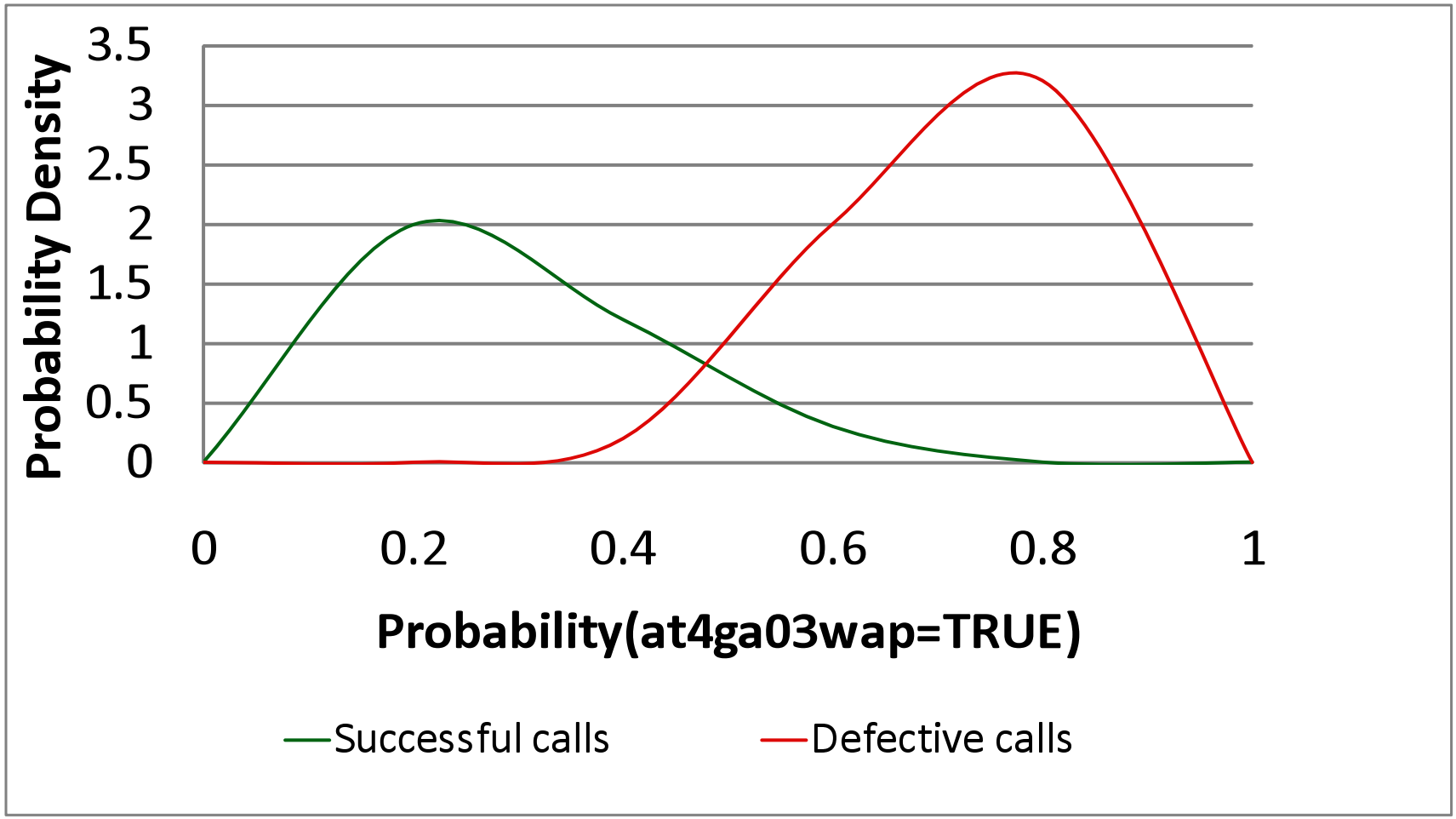
Bayesian estimation:

- Construct failure/success distributions for attribute occurrence probabilities:
  - P[Attribute occurs in failed calls],
  - P[Attribute occurs in successful call]
- Each CDR updates either failure or success distribution

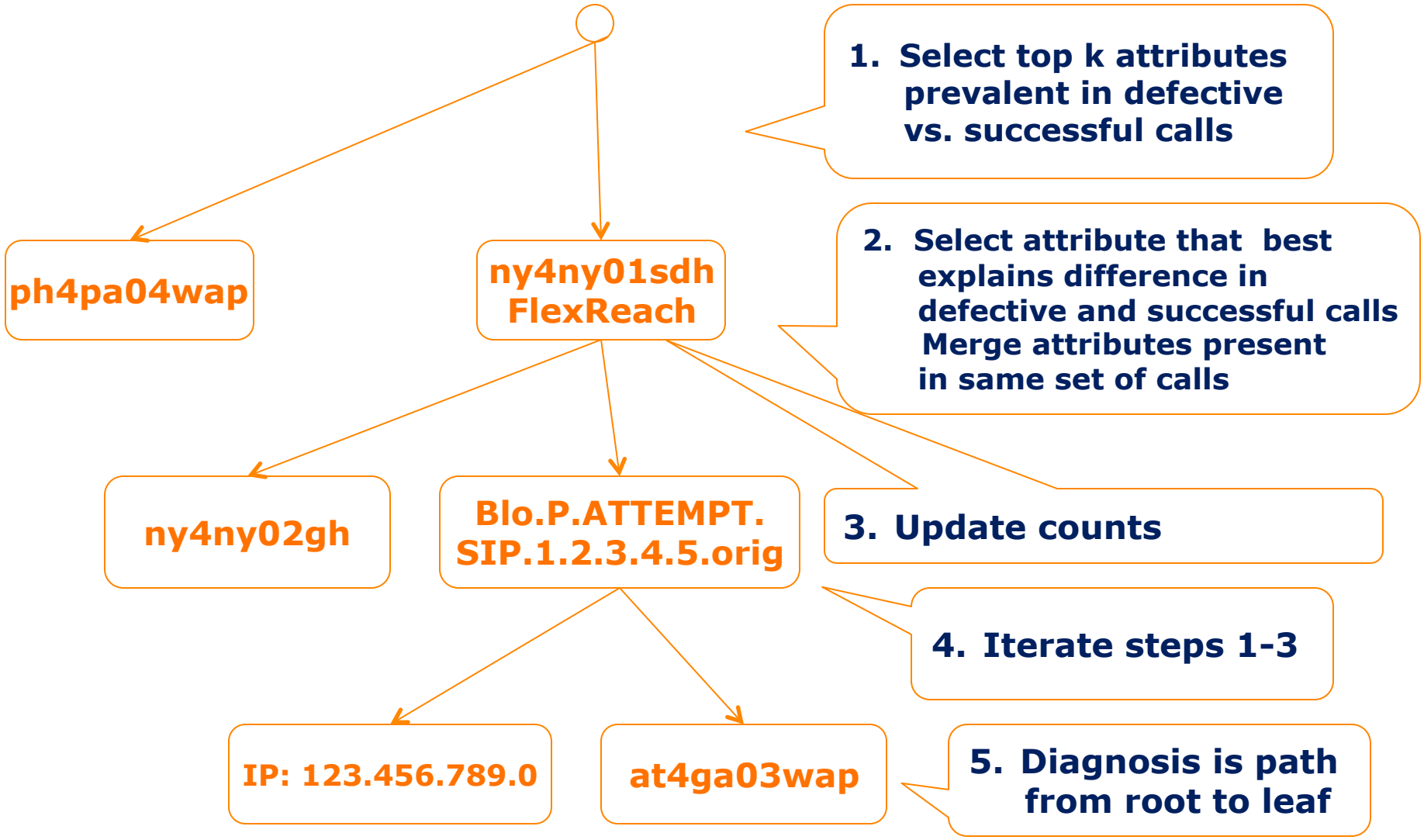
Distribution divergence

- Compute the difference between success and failure distributions
- Attribute with largest divergence is chosen as the suspect attribute

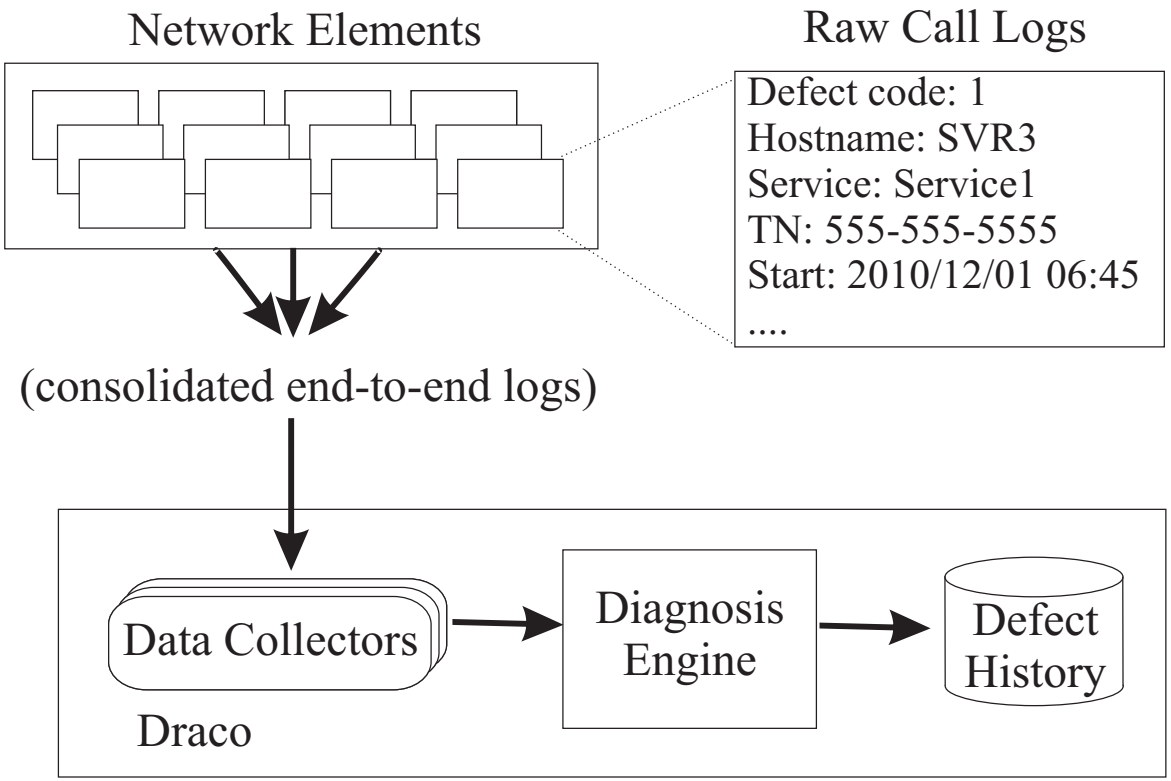
Success and Failure Attribute Distributions



Iterative Bayesian approach



Diagnosis Tool Architecture



Datasets are 10s of millions of records!

Operator Dashboard

1. Browse defects

By date

By type

- (•) Platform defects
- ( ) Non-platform defects
- ( ) All defects
- ( ) Service A
- ( ) Service B

2. View ranked list of defects

Date - 2011/1/1
Total calls: 1000000
Total defects: 1000

Defect Signature1

SRV3	Calls affected: 100
PHONE1	% of total defects: 50%

3. View samples of calls affected

⊕ Sample calls

Defect code: 1
Origin: IP
Service: Service1
TN: 555-555-5555
Start: 2010/12/01 06:45
End: 2010/12/01 06:50

4. View defect frequency for day

5. Identify chronic problems

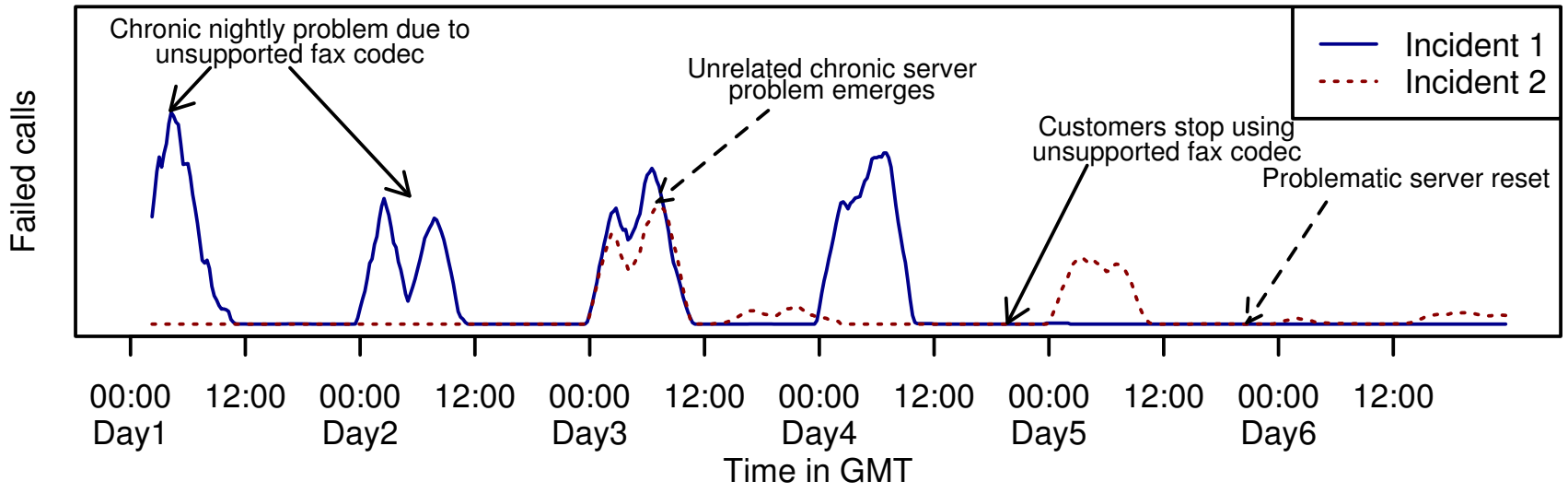
⊕ Defect history

Case Study

Used to assist in identifying root cases of *chronic defects:*

- Low impact defects that persist for days or weeks
- Six months of use

Success stories



Key take Away

A top down statistical approach that minimizes the need for domain expertise and implemented in a scalable tool

Conclusions

Lessons:

- Scale and complexity make the reality daunting!
- Examples—G-RCA and Draco—illustrate different techniques for dealing with these.
- Importance of data handling.
- Service quality as a key metric.
- Value of working with real data and real systems.

➥ **Many years of research left!**

Rethink Possible