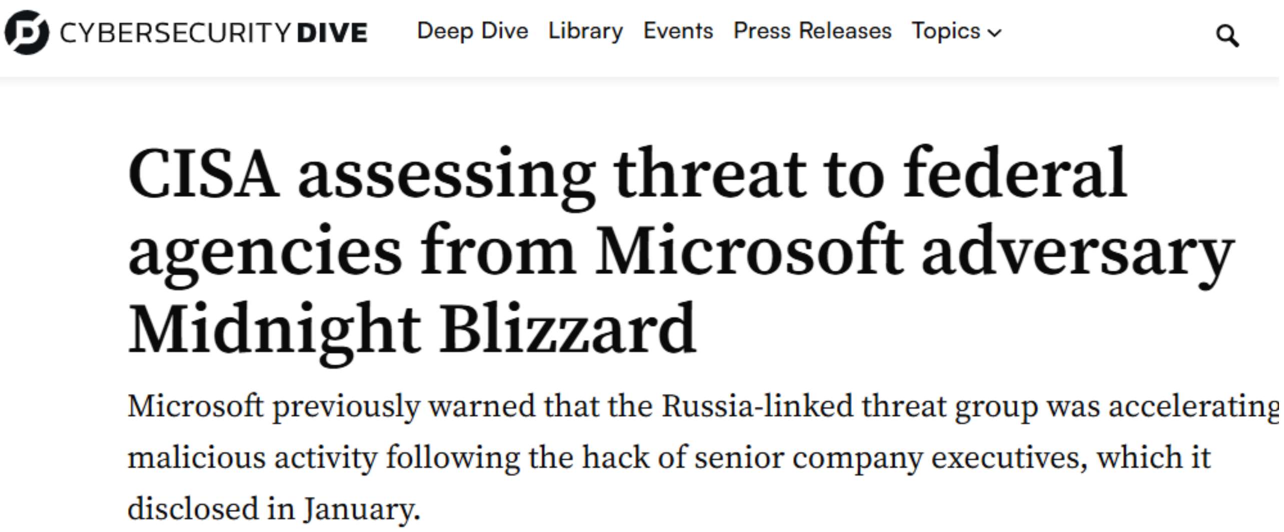
Securing Public Cloud Networks with Efficient Role-based Micro-Segmentation

Sathiya Kumaran Mani*, Kevin Hsieh*, Santiago Segarra
Ranveer Chandra, **Yajie (Lesley) Zhou**, Srikanth Kandula

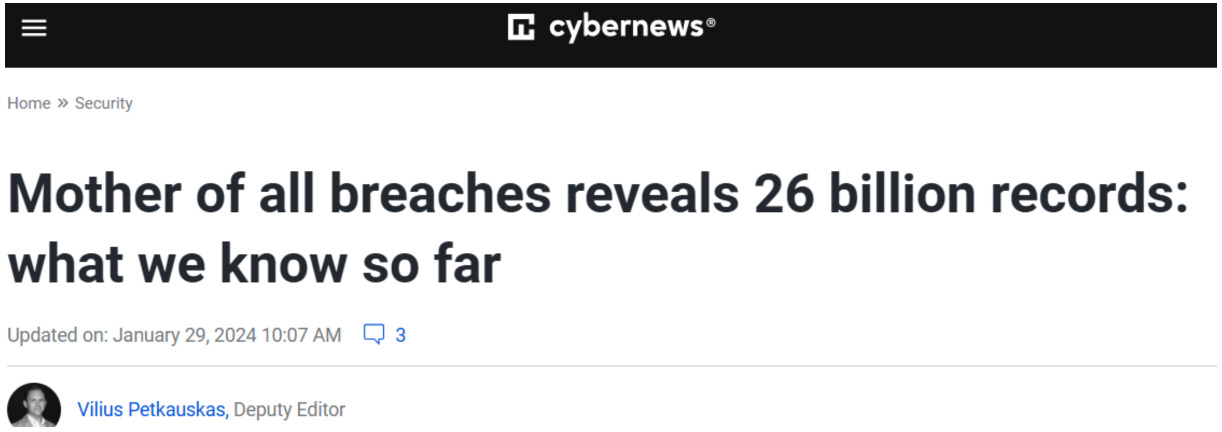
Microsoft RICE UNIVERSITY UNIVERSITY OF MARYLAND

Public Clouds are Major Targets for Cybercrimes



CISA assessing threat to federal agencies from Microsoft adversary Midnight Blizzard

Microsoft previously warned that the Russia-linked threat group was accelerating malicious activity following the hack of senior company executives, which it disclosed in January.

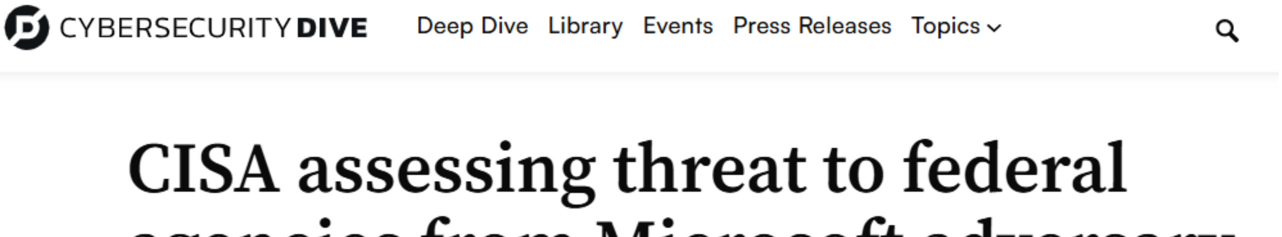


Mother of all breaches reveals 26 billion records: what we know so far

Updated on: January 29, 2024 10:07 AM 3

Vilius Petkauskas, Deputy Editor

Public Clouds are Major Targets for Cybercrimes

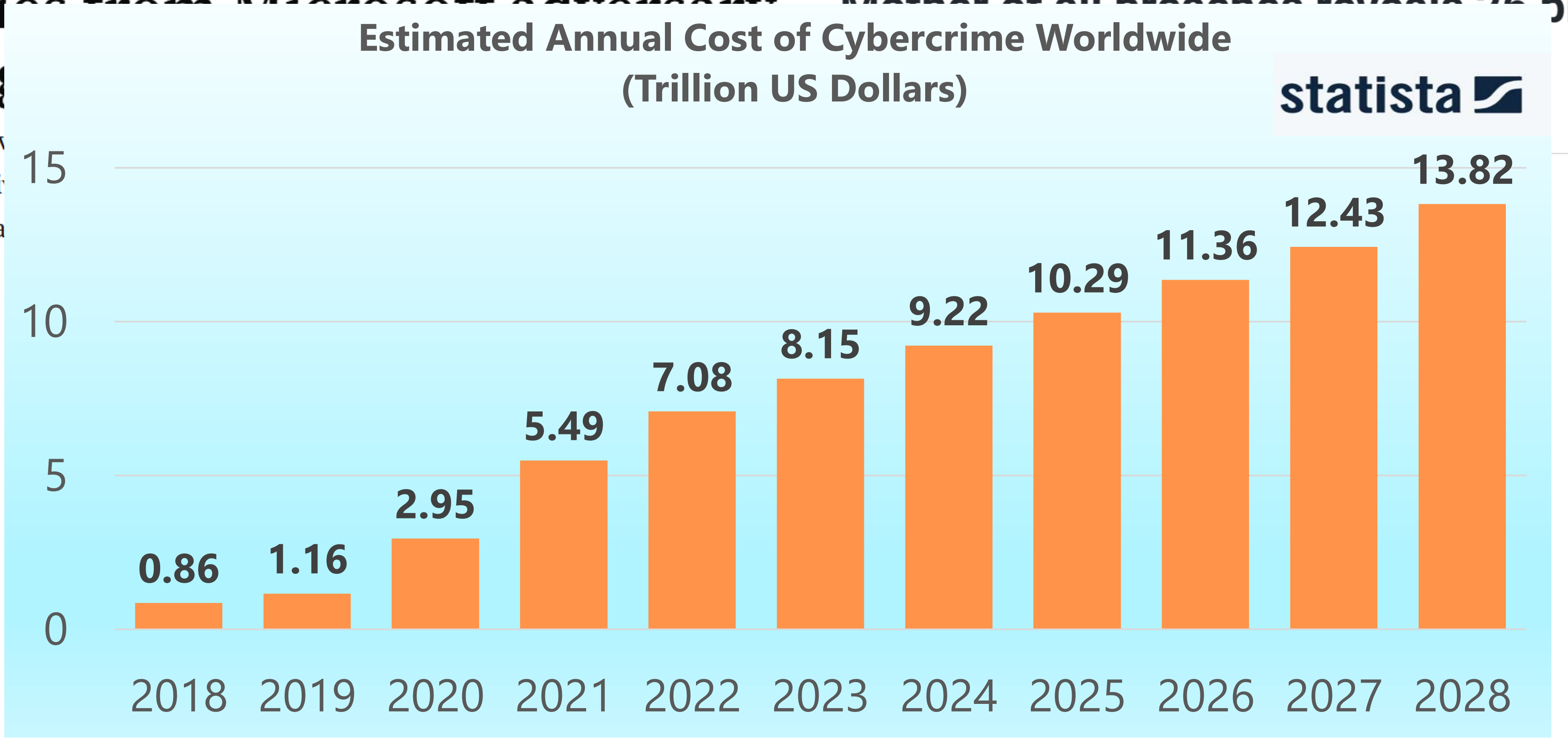


CISA assessing threat to federal agenci from Microsoft adversary Midni

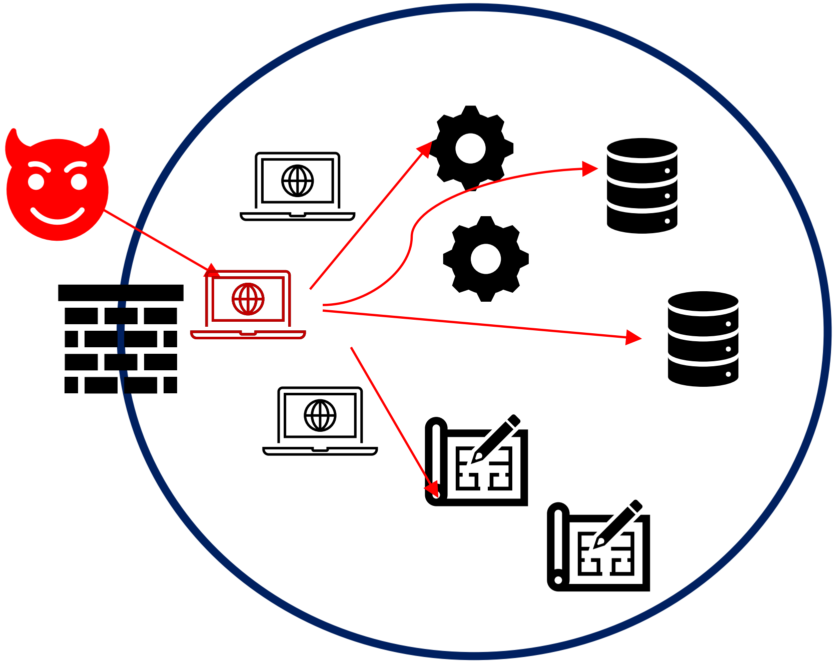
Microsoft prev
malicious acti
disclosed in Ja

Home » Security

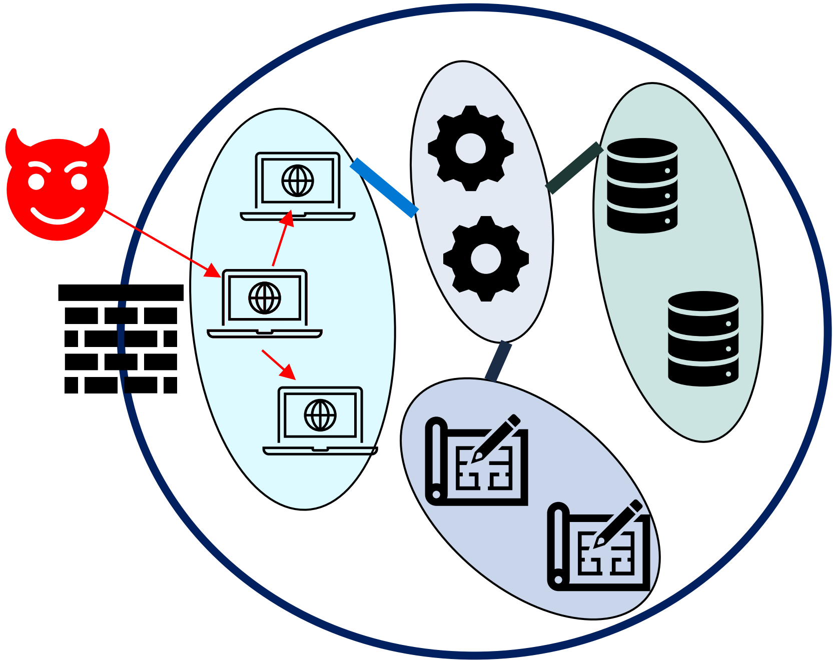
Mother of all breaches reveals 26 billion records:


Estimated Annual Cost of Cybercrime Worldwide (Trillion US Dollars)

Promising Solution: Micro-Segmentation

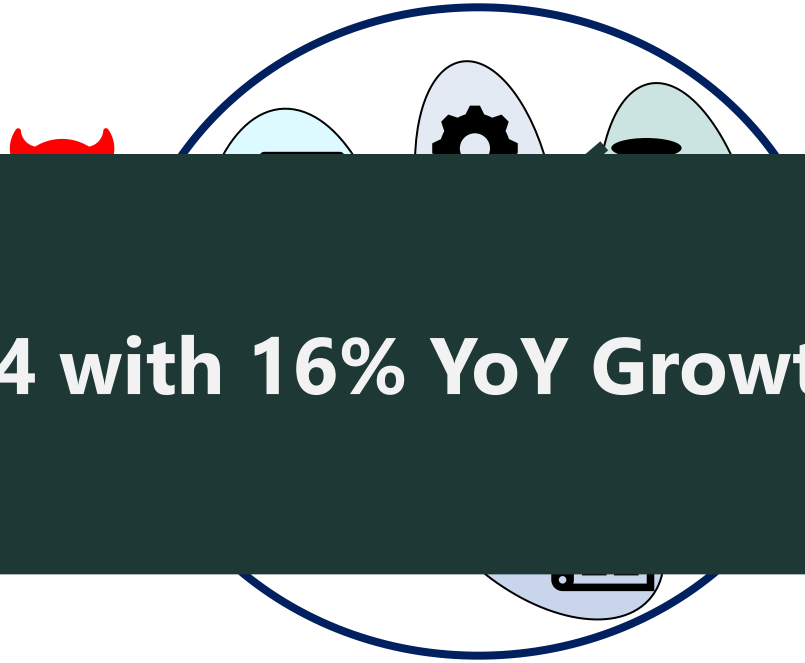


Ring Fencing



Ring Fencing + Micro-Segmentation
(Zero Trust Architecture)

Promising Solution: Micro-Segmentation



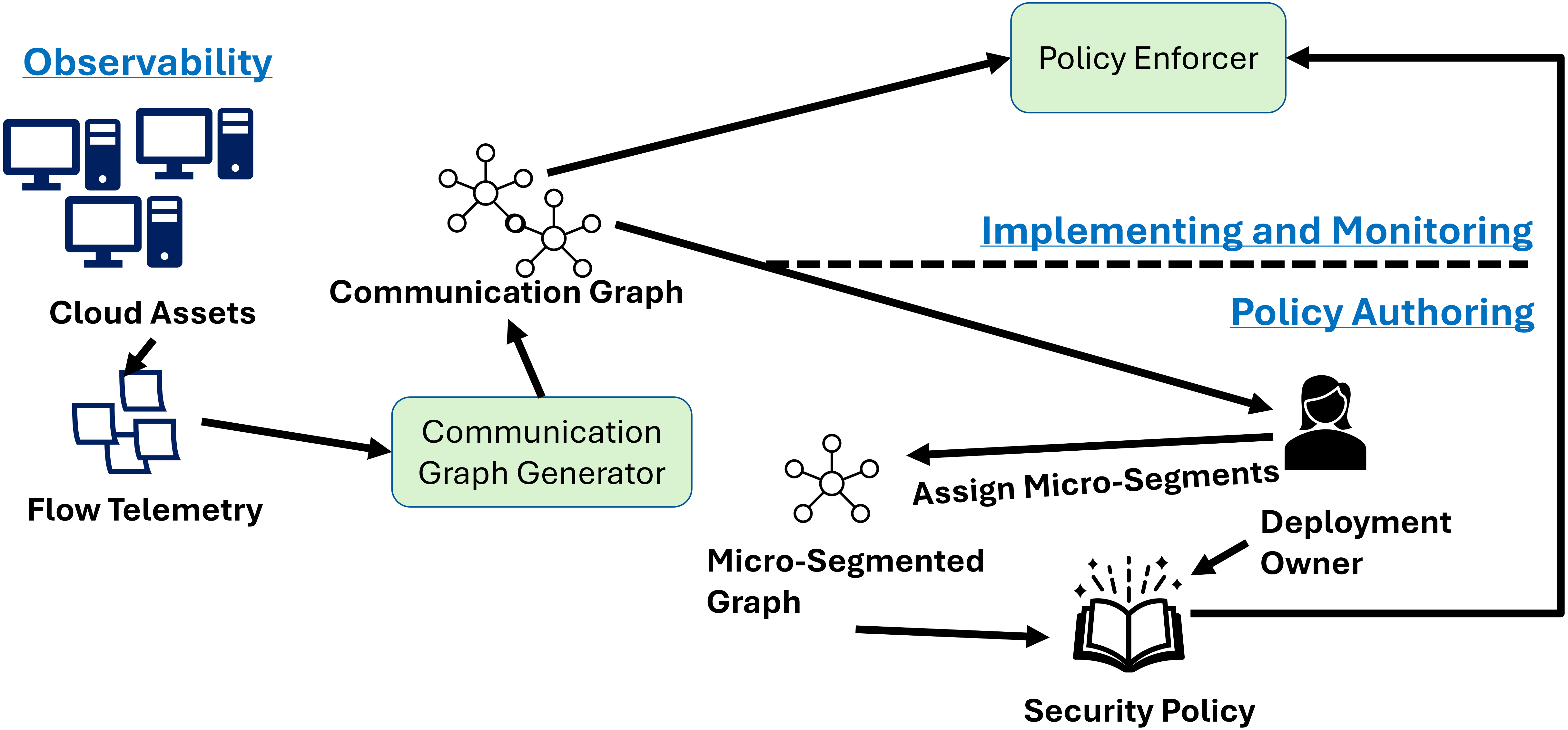
Total Market $20B in 2024 with 16% YoY Growth

Ring Fencing

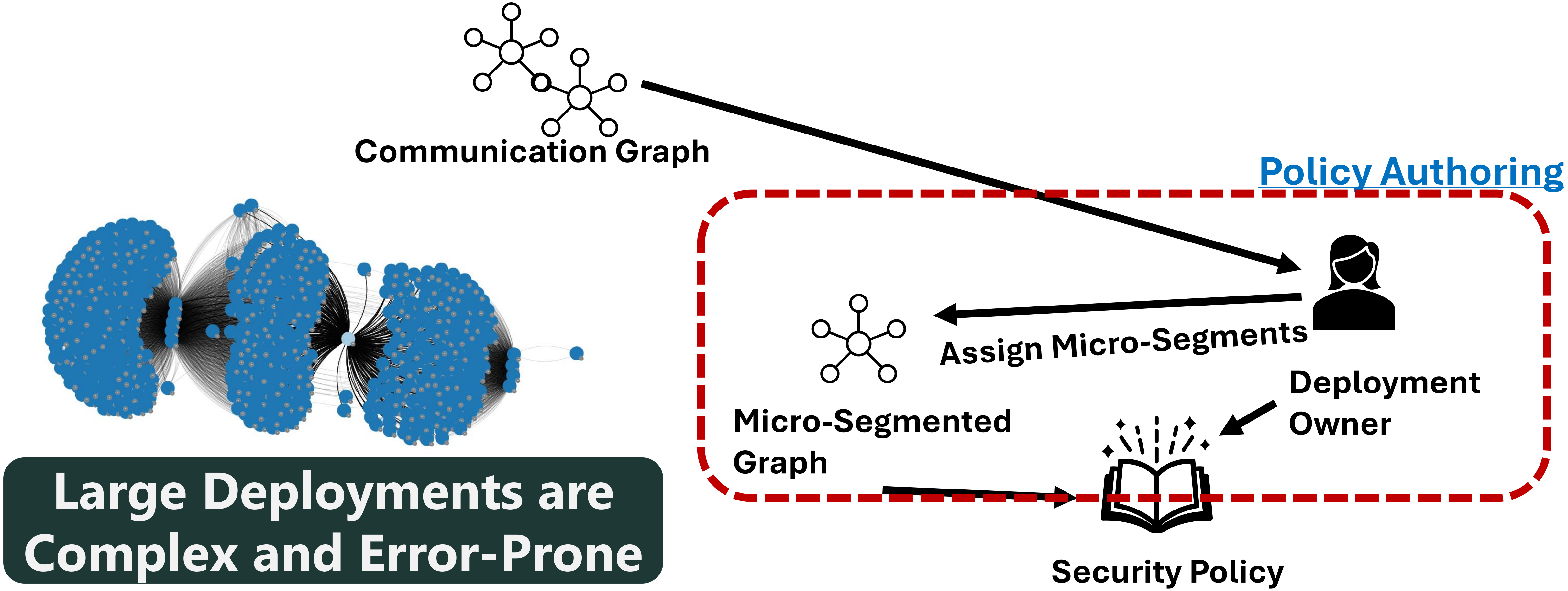
Ring Fencing + Micro-Segmentation
(Zero Trust Architecture)

[https://www.researchandmarkets.com/report/microsegmentation]

Micro-Segmentation Workflow



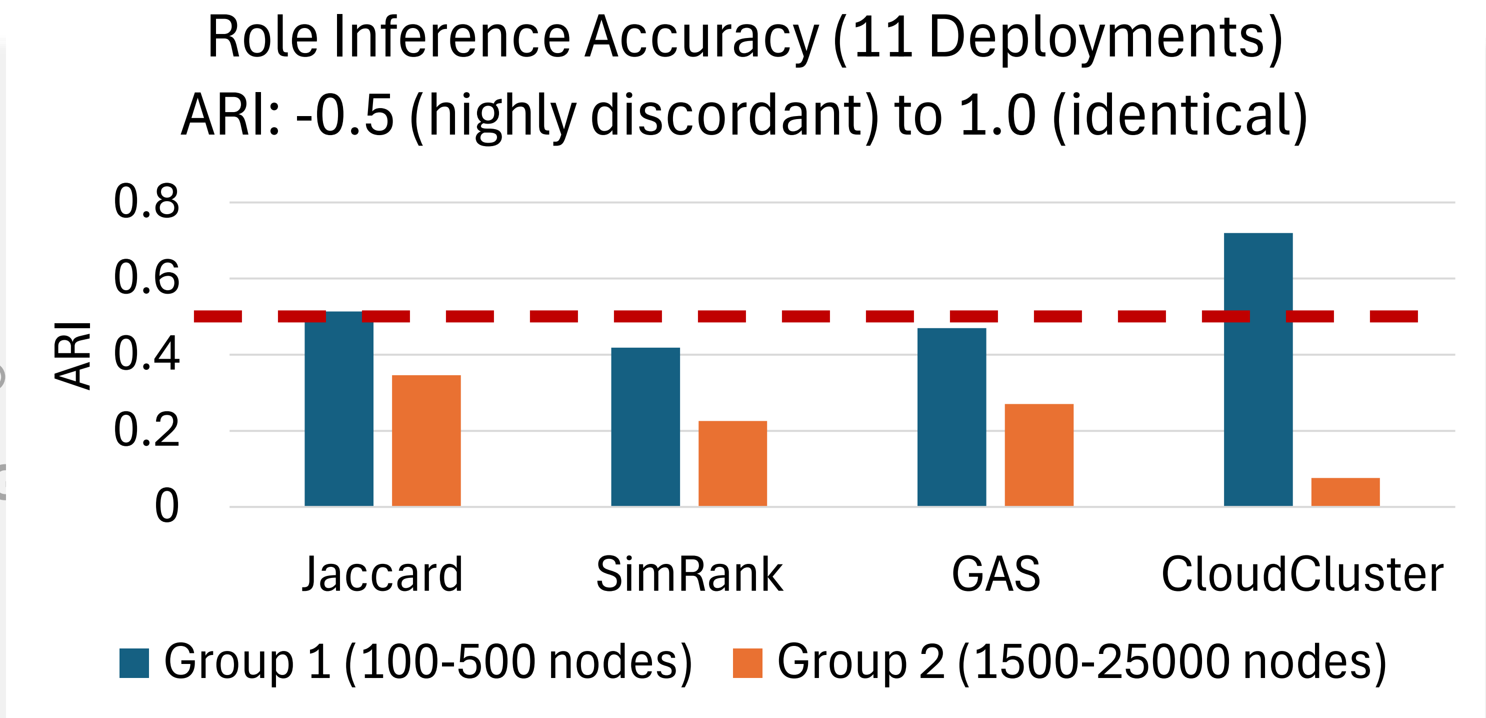
Challenge 1: Manual Micro-Segments Assignments



Large Deployments are Complex and Error-Prone

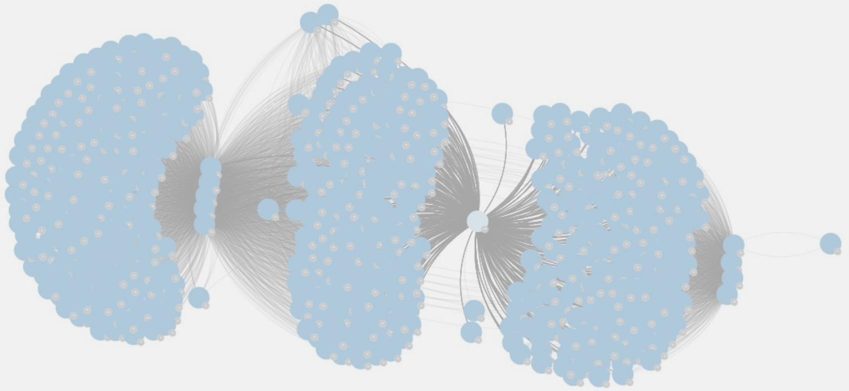
Challenge 1: Manual Micro-Segments Assignments

Can Role Inference Algorithms Help?


Role Inference Accuracy (11 Deployments)

ARI: -0.5 (highly discordant) to 1.0 (identical)

Communication G



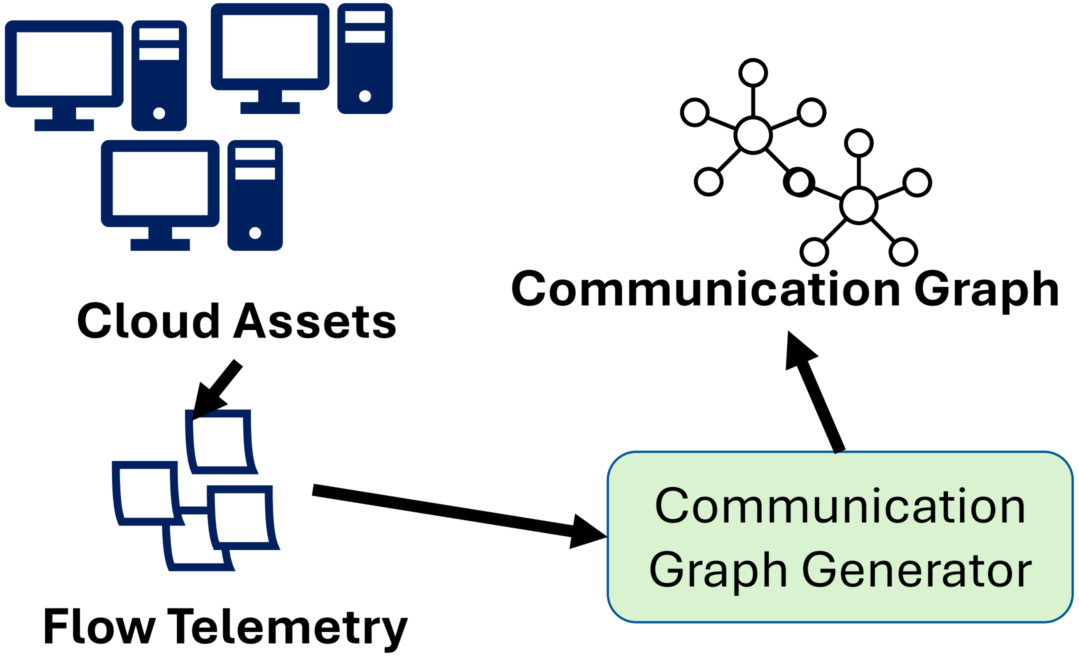
Large Deployments are Complex and Error-Prone

Existing Role Inference Algorithms Are Very Inaccurate

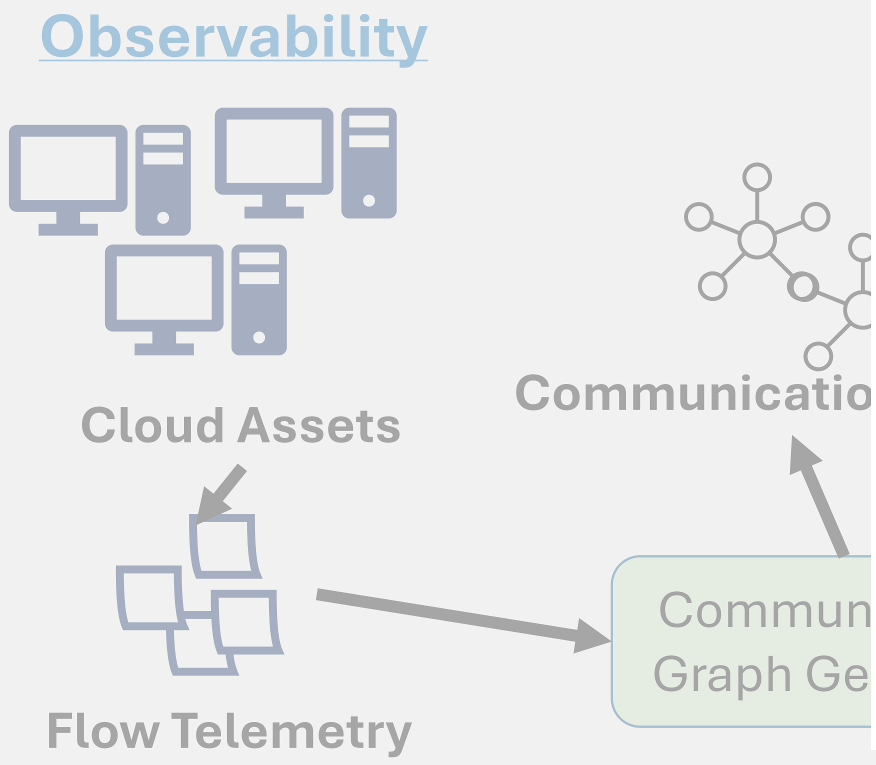
Security Policy

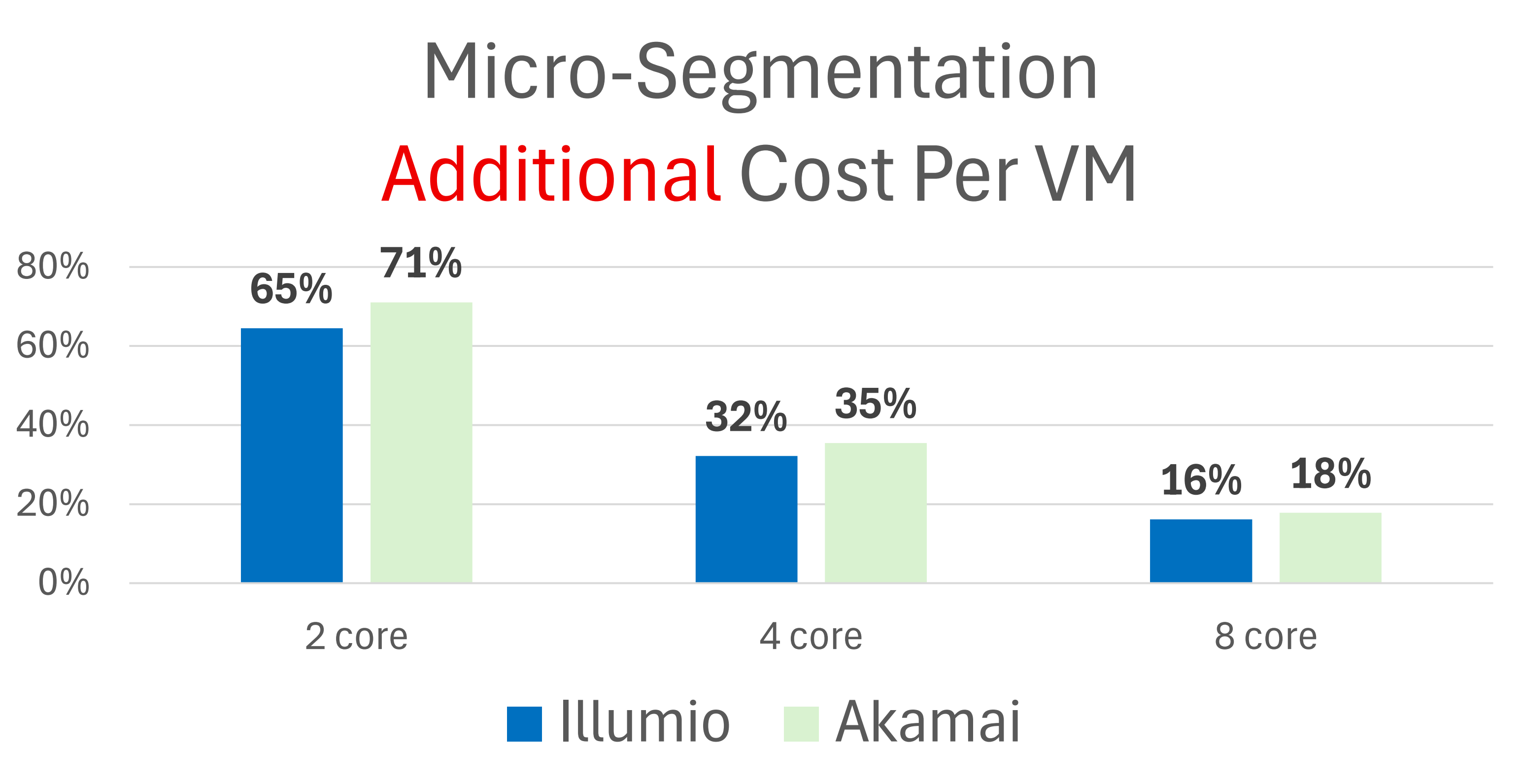
Challenge 2: Graph Generation is Costly

Observability



Challenge 2: Graph Generation is Costly




Micro-Segmentation
Additional Cost Per VM

High-Cost Overheads Hinders Widespread Adoption

Our Solution: ZTS (Zero Trust Segmentation)

Role-Inference Algorithm for Micro-Segmentation

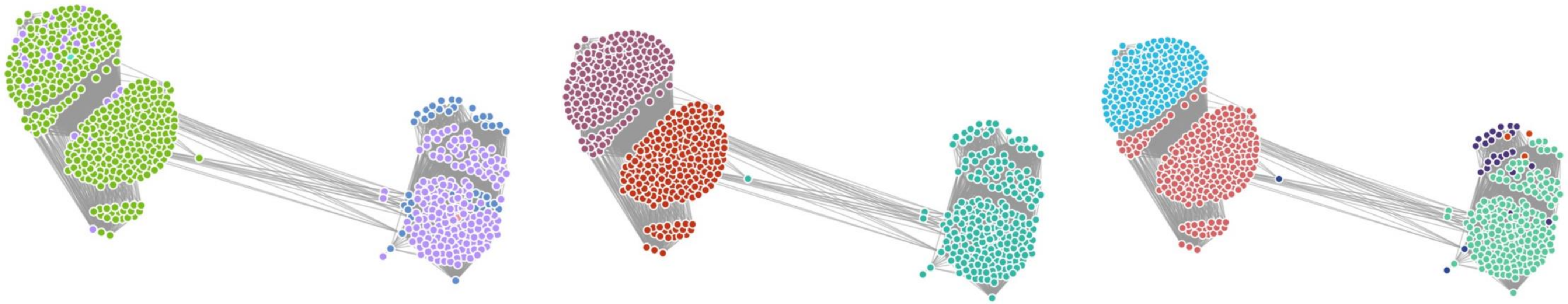
- Facilitate the creation of precise, scalable security policies

Cost-Effective Communication Graph Generator

- A scalable and low-cost architecture to generate communication graphs

Existing algorithms are insufficient on production graph

All produce very different results – Far from the ground truth

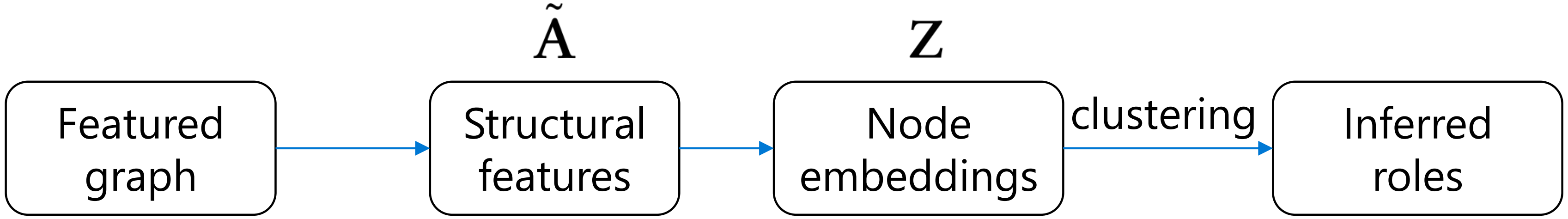


(a) Simrank segmentation

(c) Conn.-weight. modularity

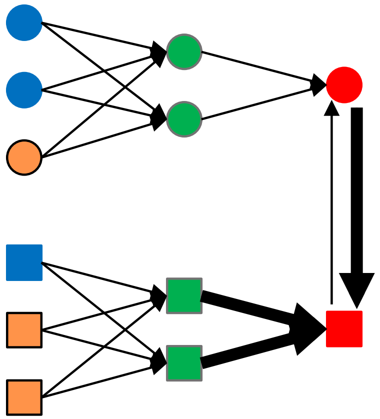
(d) Byte-weighted modularity

Existing role-inference approach all based on graph structural features



$\tilde{A}$ Z

Featured graph → Structural features → Node embeddings → clustering → Inferred roles

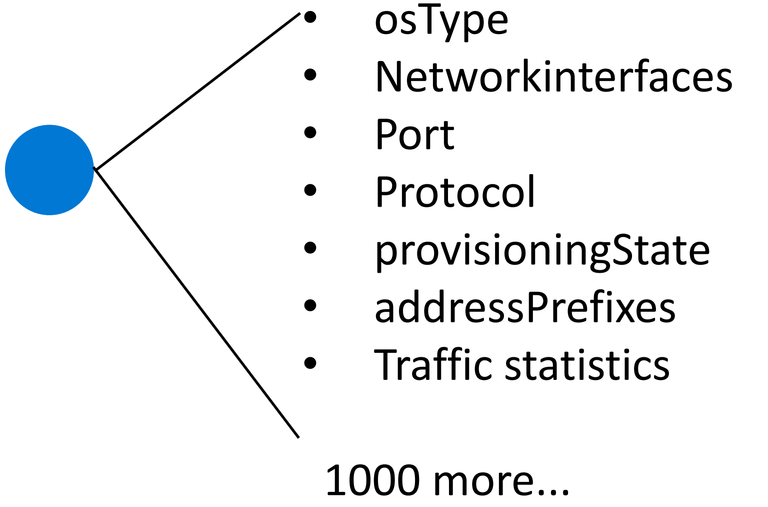


Roles only based on color

Graph structure can look identical, but we have a lot more info

Challenges: Not trivial to feed domain knowledge...

Which features are the most important?



- osType
- Networkinterfaces
- Port
- Protocol
- provisioningState
- addressPrefixes
- Traffic statistics

1000 more...

Feature importance changes

Deployment A:

- addressPrefixes
- Port
- Protocol

Deployment B

- osType
- Networkinterfaces
- provisioningState

Opportunity: there exists sparse labels!
How can we use it to help us infer roles?

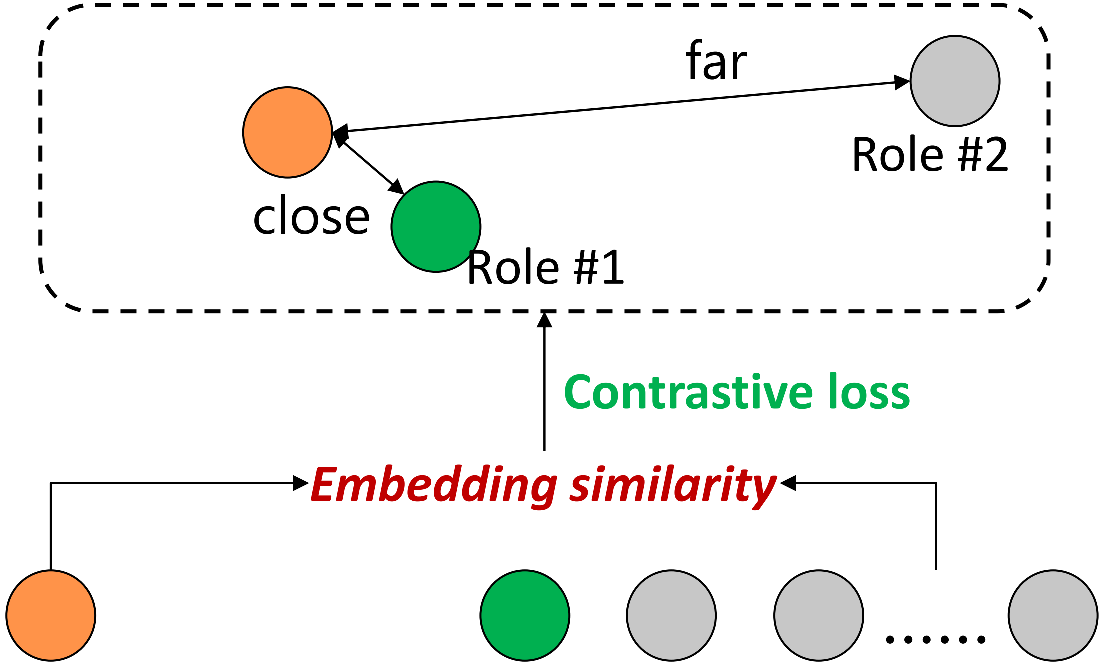
🤔

Opportunity: Contrastive learning

Intuition: pull embeddings of similar pairs together, pushes dissimilar pairs apart.

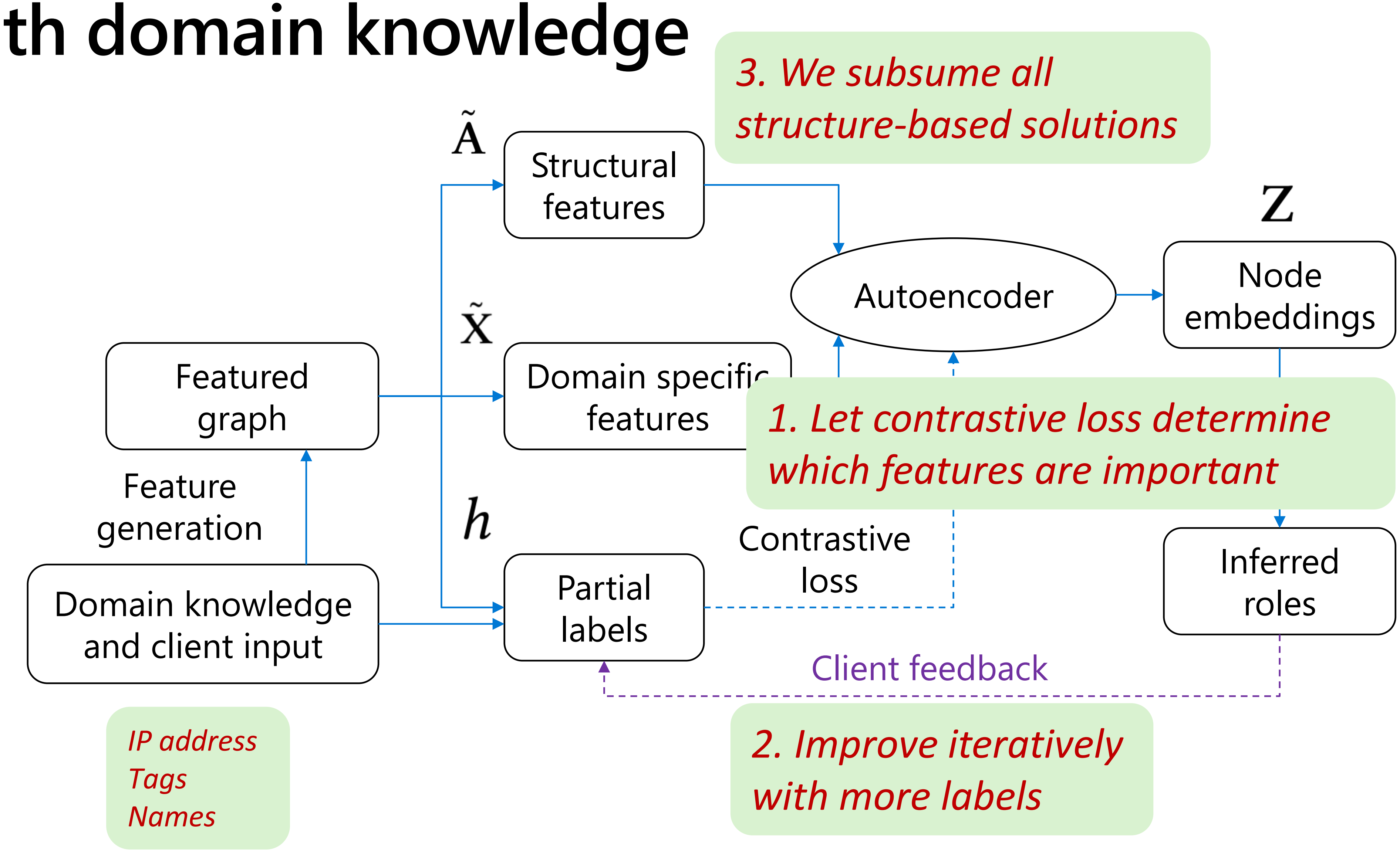
$$L(\theta,\psi) = \sum_{i\in\mathcal{V}} \|\mathbf{y}_i - g_\psi(f_\theta(\mathbf{y}_i))\|_2 -$$

$$\alpha \sum_{r\in\mathcal{R}} \sum_{\substack{i,i'\in\mathcal{L}\\ h(i)=h(i')=r}} \log\left(\frac{\exp(\mathrm{sim}(f_\theta(\mathbf{y}_i), f_\theta(\mathbf{y}_{i'}))/\tau)}{\sum_{\substack{i''\in\mathcal{L}\\ h(i'')\neq r}} \exp(\mathrm{sim}(f_\theta(\mathbf{y}_i), f_\theta(\mathbf{y}_{i''}))/\tau)}\right).$$



Our idea: Use partial labels to refine and guide the role-inference

New role-inference algorithm: contrastive learning with domain knowledge



3. We subsume all structure-based solutions

1. Let contrastive loss determine which features are important

2. Improve iteratively with more labels

IP address
Tags
Names

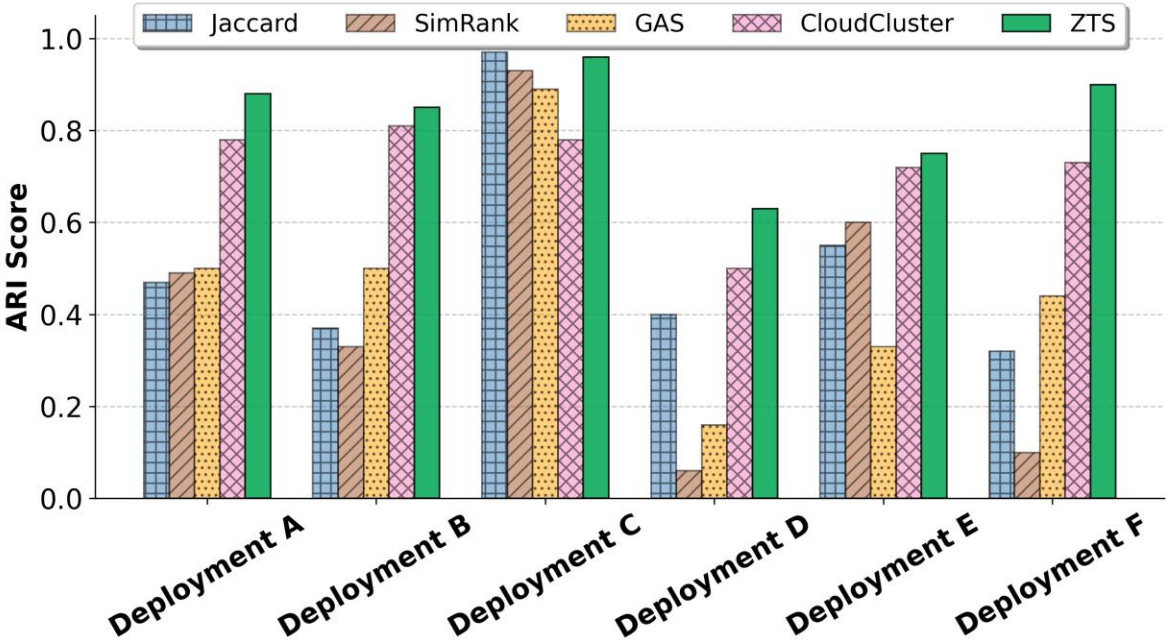
Role Inference Results: ZTS is consistently the best

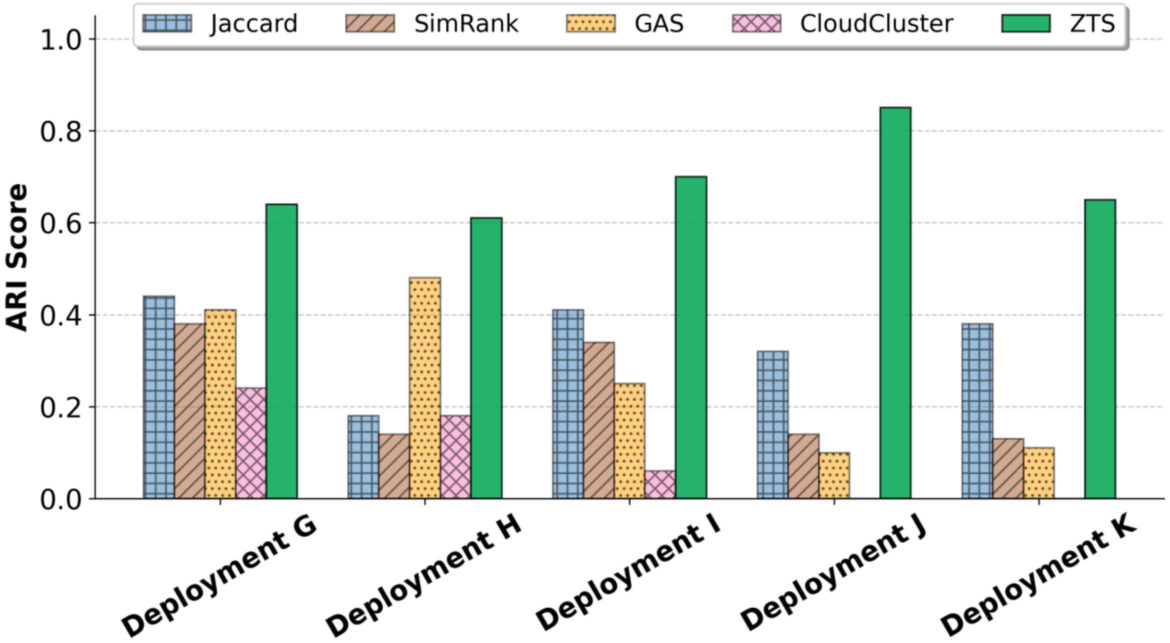
Small deployments

- Node: 100-200
- Edges: 100-9000
- Roles: 12-28

Large deployments

- Node: 1500-25000
- Edges: 5000-165000
- Roles: 20-87





On Average, ZTS: 0.77 vs Best Baseline: 0.43

Our Solution: ZTS (Zero Trust Segmentation)

Micro-Segmentation with Role Inference

- Facilitate the creation of precise, scalable security policies

Cost-Effective Communication Graph Generator

- A scalable and low-cost architecture to generate communication graphs

Building a system for Graph Generation

Goal:

Use systems available in large public clouds to be cost-effective and scalable

Low cost crucial for extensive adoption

Telemetry source: Network flow (or connection) summaries

Cost-effective
Tamper-proof
Gathered with minimal disruption

Building a system for Graph Generation

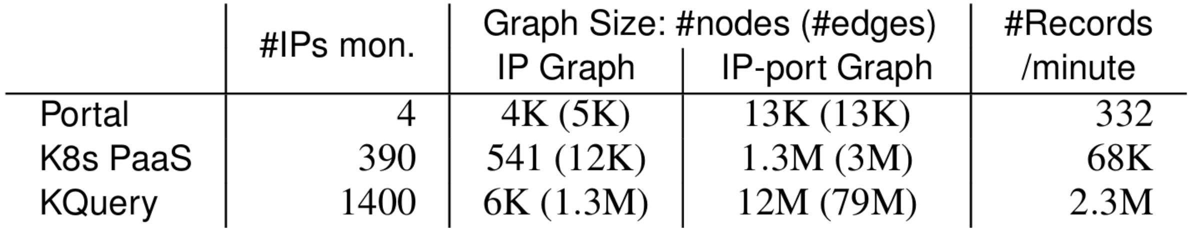
Practical challenges:

Structure of telemetry

Numerous small files (one JSON file/hour/VM)

Volume of telemetry

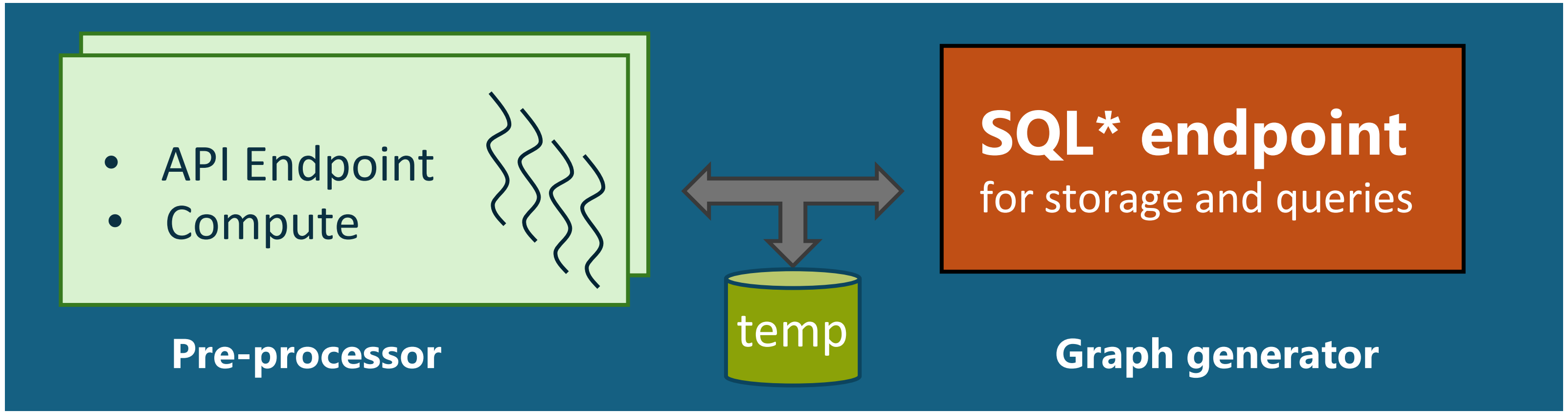
#Records/min can be large
Resulting communication graph can be very large



	#IPs mon.	Graph Size: #nodes (#edges)		#Records
		IP Graph	IP-port Graph	/minute
Portal	4	4K (5K)	13K (13K)	332
K8s PaaS	390	541 (12K)	1.3M (3M)	68K
KQuery	1400	6K (1.3M)	12M (79M)	2.3M

Careful considerations needed to keep processing time and cost low

Building a system for Graph Generation



- Allows us to pipeline the stages and parallelize pre-processing → Handle telemetry volume
- Optimize data format for large SQL batch processing → Address telemetry structure

We further optimize the SQL query processing,

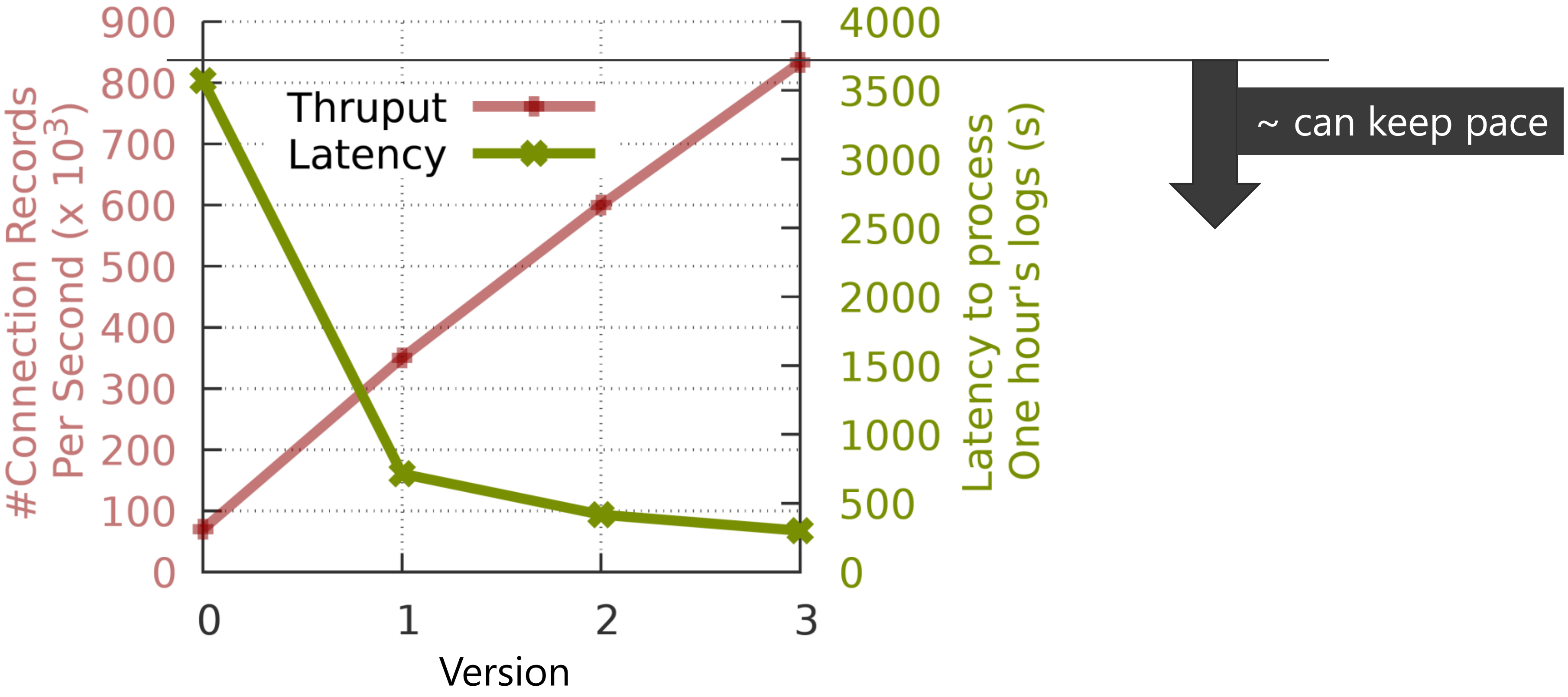
Avoid naïve group-by-aggregation – focus on heavy hitters
Utilize Common Table Expressions (CTEs) – optimal query plan, avoid materialization
…

Improvements with Designs and Optimizations

Deployment with
700 VMs
~ 250 M records
per hour

Using one VM (64c)
and two server SQL
processing instances

IP-graph with 0.1%
contribution cutoff



~ can keep pace

Can run at low-cost: a surcharge of 0.05% (e.g., 3 boxes for ~ 5000 VMs)

Comparative Experiments

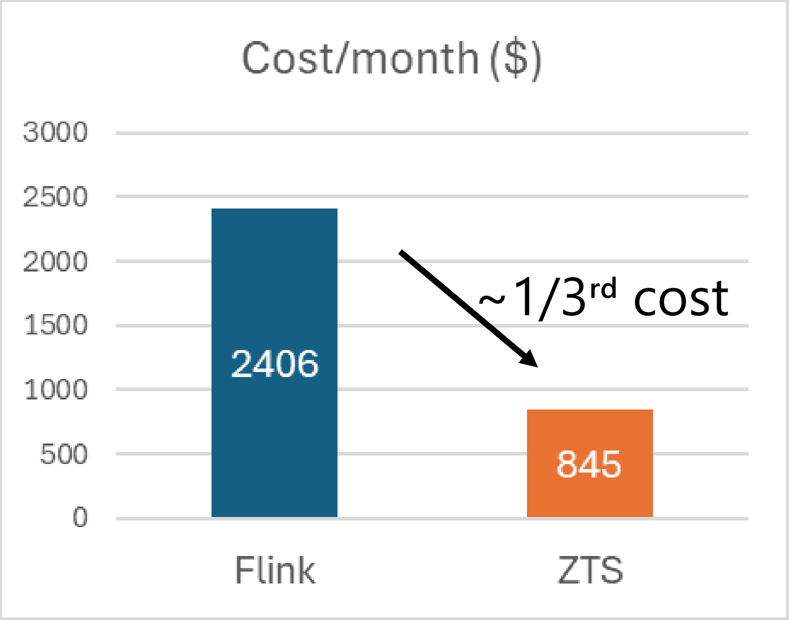
We built a streaming pipeline (based on OSS Apache Flink)

Enterprise-ready
Strong customer support
Highly performant
Low resource consumption

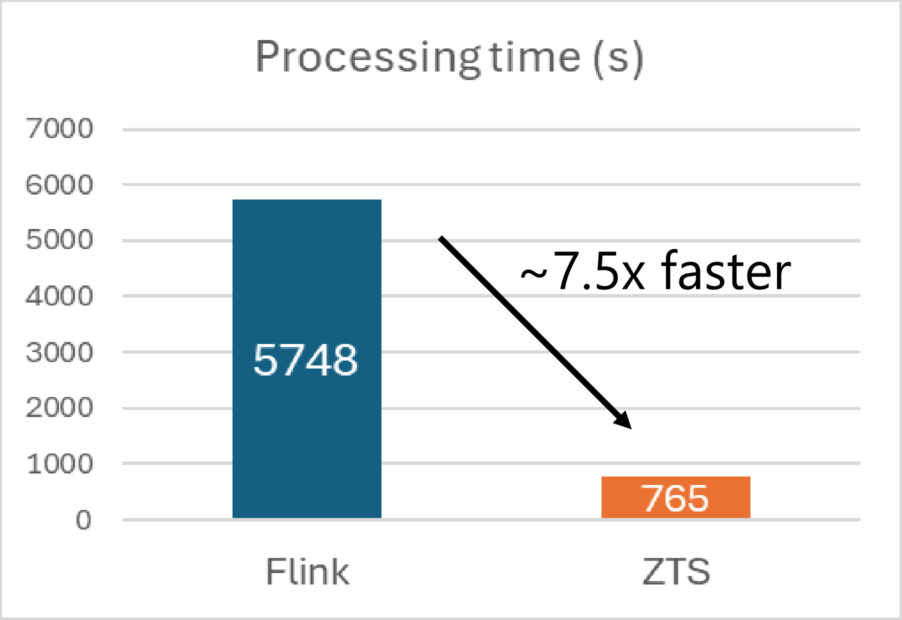
Resources

- ZTS: 1 VM (8 cores + 32G) + 1 server SQL instance
- Flink: 1 VM (64 cores + 256G)

At scale,
(39M recs/hr)



~1/3rd cost



~7.5x faster

ZTS is **21x more cost effective**

Conclusion

Implementing micro-segmentation at scale remains challenging

ZTS is a novel end-to-end system,

- Effective role-inference algorithm to facilitate security policy authoring
- Scalable network communication graph generation

Using real-world deployments we show,

- The performance of contrastive learning with domain knowledge
- Cost effectiveness of our system implementation

Thank you!