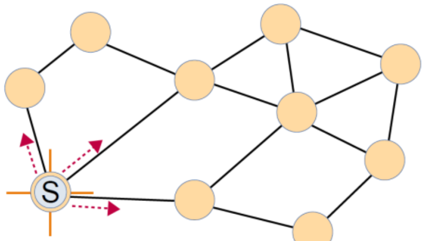
Workload Diffusion Modeling for Distributed Applications in Fog/Edge Computing Environments

The International Conference on Performance Engineering

Thang Le Duc
thang@cs.umu.se, thang.leduc@tieto.com

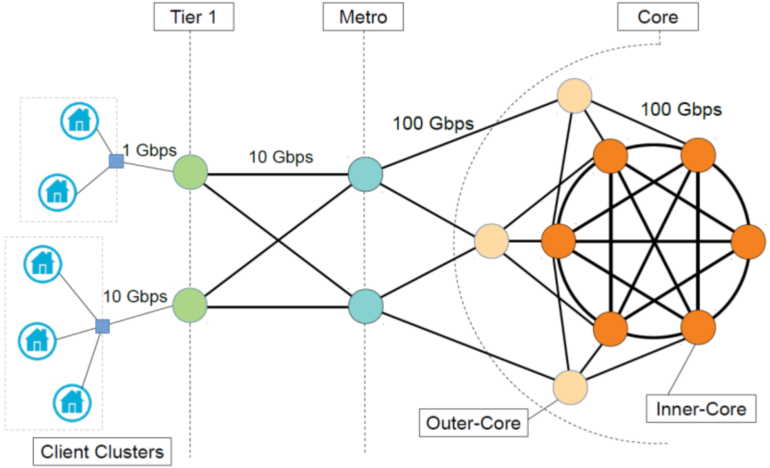
2020 - 04 - 24

Motivation



- Peer-to-peer overlay network
- Ad hoc network

- Limitation in workload data collection in large-scale distributed applications/systems [1]
- Workload propagation model [1,2]
  - How workload from a node is propagated to its neighbors?
  - To facilitate workload preditions and/or workload generation
    - Auto-scaling and system remediation (in RECAP: https://recap-project.eu/)



- Content Delivery Network (CDN)
- Core Broadband Network

Agenda

- Introduction
- Non-Hierarchical Workload Diffusion
- Hierarchical Workload Diffusion
- Experiments
- Discussion
- Conclusions

Introduction

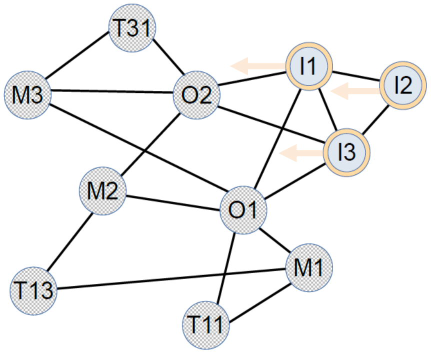
Issues and Challenges

- The necessity of understanding the applications and their workload behaviors
  - Large-scale distributed applications in fog/edge computing environments: CDN, telco network services, IoT application, …
  - Workload and/or application characterization, analysis and modeling
  - Workload propagation models
- The high demand of publicly available datasets
  - Time series datasets: web traffic, system resource utilization, …
  - Synthetic workload generation for diverse applications

tieto

Introduction
Problem and Solution

- Problem (*see the figure*)
  - Given workload measurements at a limited subset of nodes, generate/extrapolate supplementary workloads for the entire application/network

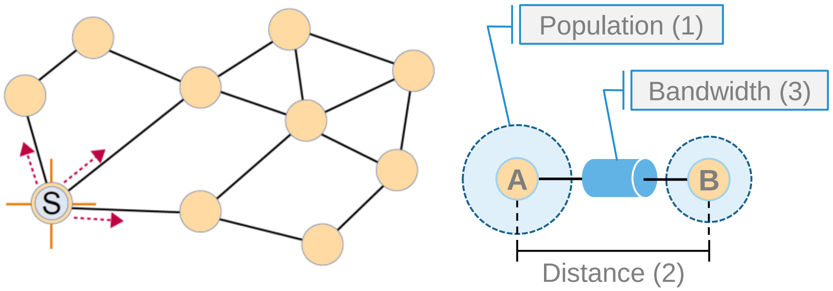


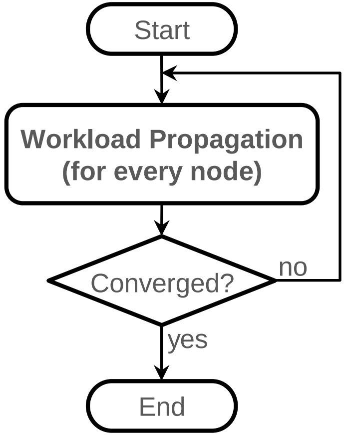
- Solution
  - Application models and/or workload propagation models
  - Workload diffusion algorithms
    - Non-hierarchical Workload Diffusion
      - Applicable to non-hierarchical systems: unstructured peer-to-peer overlay or ad-hoc networks
    - Hierarchical Workload Diffusion
      - Applicable to hierarchical systems: CDNs or core broadband networks
  - Final target: a framework with the models and algorithms integrated

Non-Hierarchical Workload Diffusion

- Population-based Diffusion
  - Based on the population associated to nodes (1)
    - A node with larger population receives higher load from a source
- Location-based Diffusion
  - Based on the geographical location of nodes or distance between nodes (2)
    - A node closer to the source receives higher load
  - Executed in iterations as shown in the flow chart
    - Convergence: predefined threshold or no significant changes
- Bandwidth-based Diffusion
  - Based on the bandwidth capacity of links (3)
    - Workload distributed on a link is proportional to the link's capacity
  - Executed in iterations as shown in the flow chart

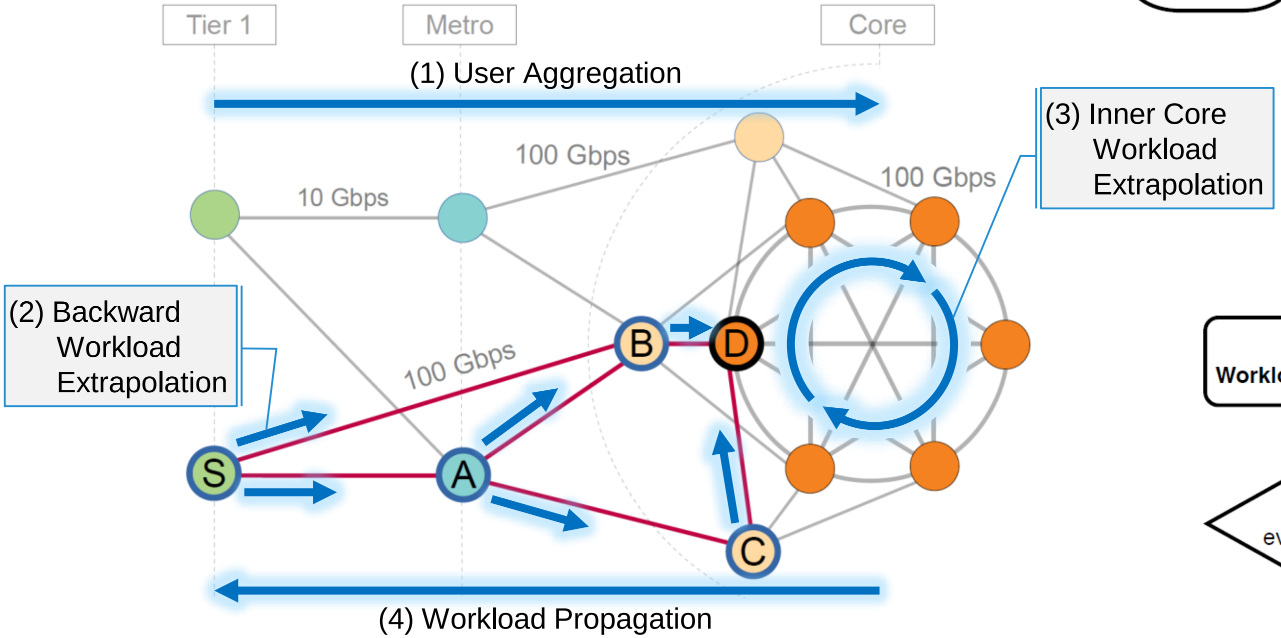
Network model

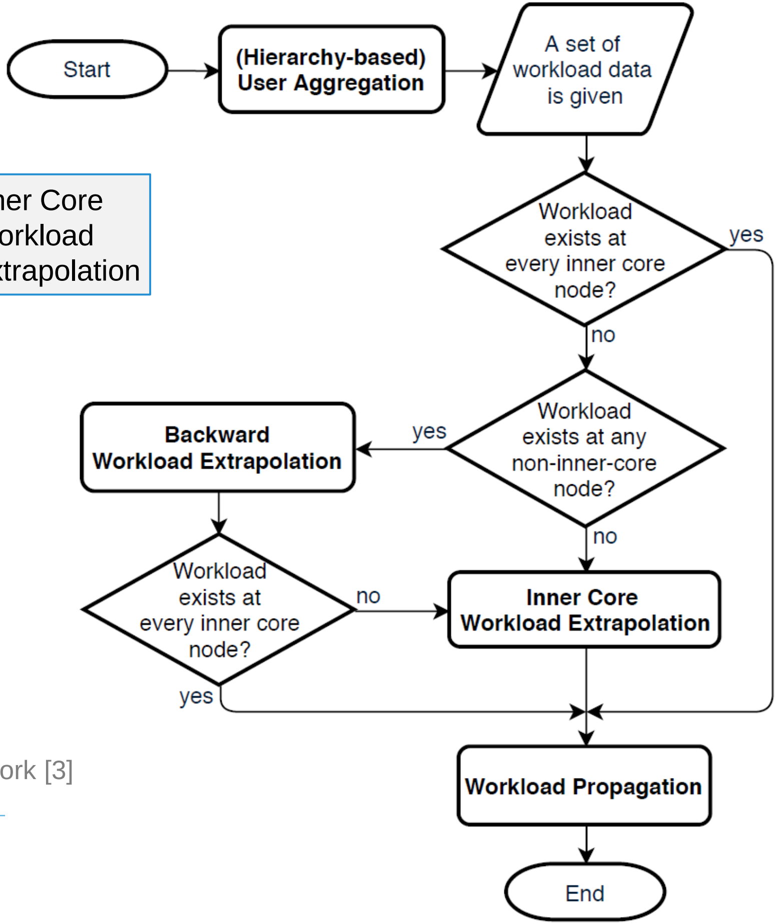




Hierarchical Workload Diffusion

Hierarchy-based Diffusion

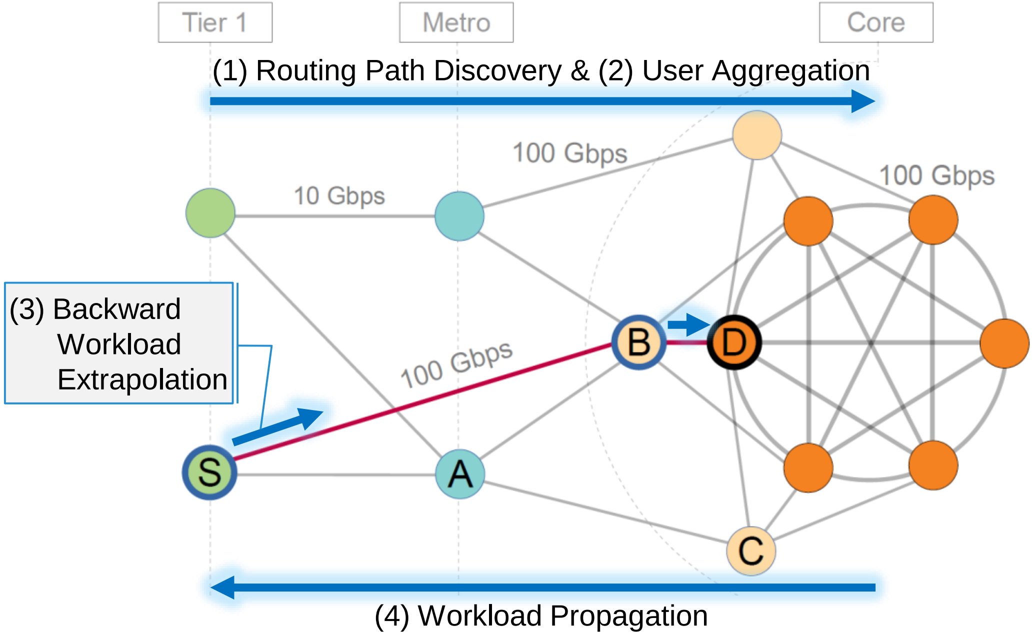


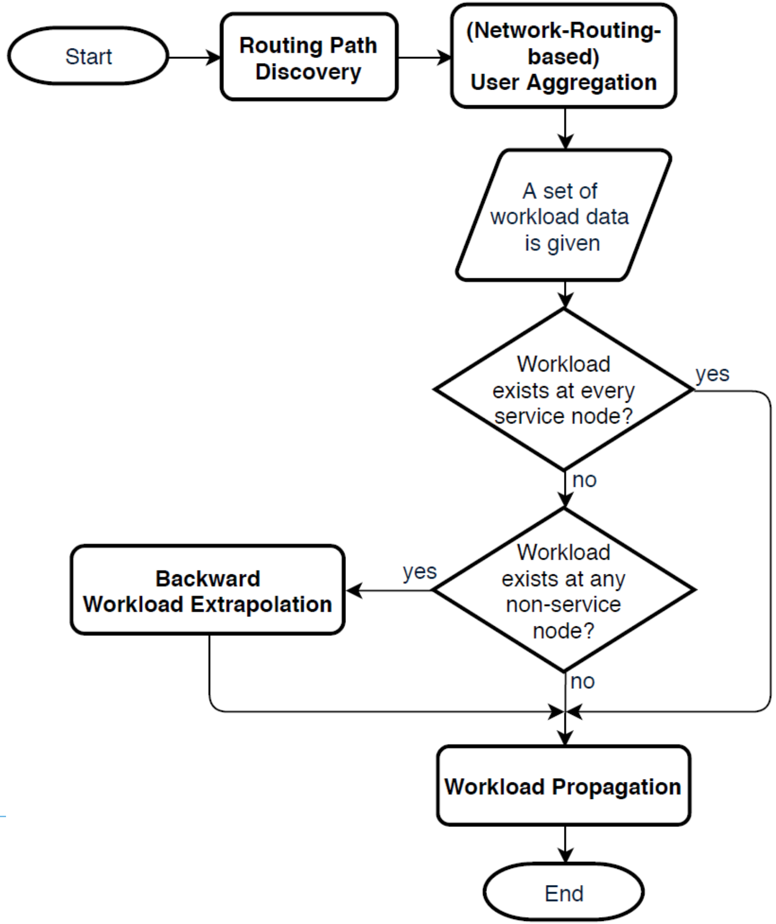


- The adopted hierarchical network model is a representative of the BT 21CN production network [3] but at a smaller scale (https://kitz.co.uk/adsl/21cn_network.htm)

Hierarchical Workload Diffusion

Network-Routing-based Diffusion





Summary

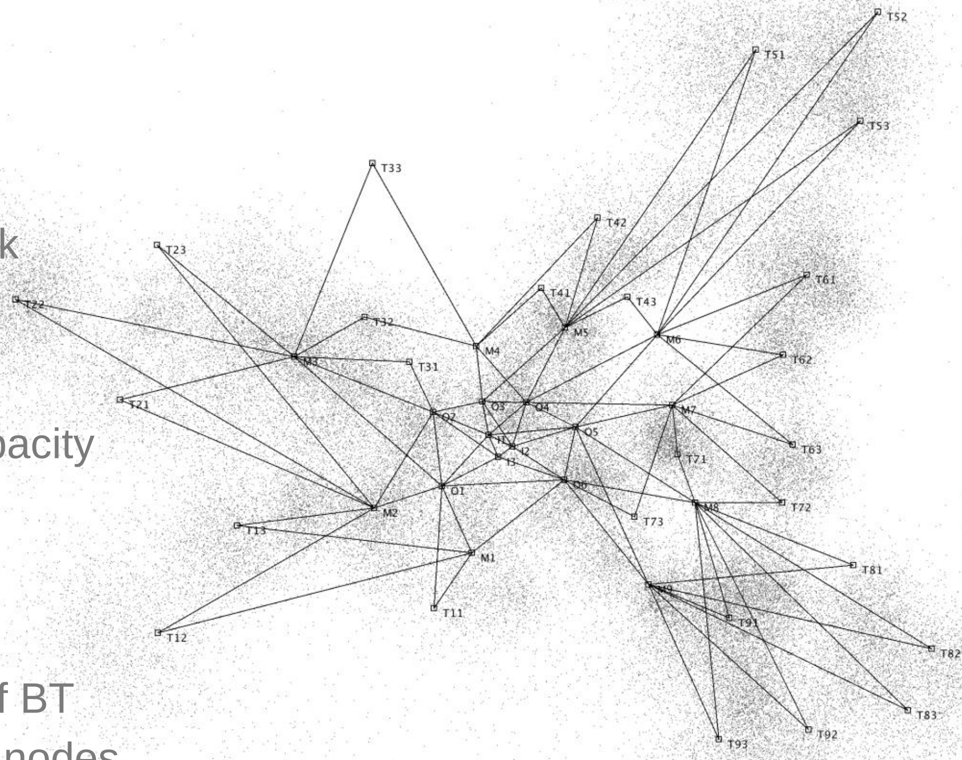
Algorithm	Assumptions	Key Inputs	Description
Population-based	• Non-hierarchical network/application topologies • Homogeneous user behavior	• User distribution in the network	• Iterative refinement algorithms (similar to heat diffusion and spring relaxation equations) • Repeatedly solve state equations to distribute workload to neighbours until the overall load distribution approaches equilibrium • Algorithms are highly parallelizable
Location-based		• Geographical node locations	
Bandwidth-based		• Bandwidth capacity of links	
Hierarchy-based	• Hierarchical network/ application topologies • Full mesh network of the inner-core nodes • Multiple shortest path routing • Homogeneous user behavior	• Network hierarchy • Bandwidth capacity of links • User distribution in the network	• User aggregation: identifies the aggregated number of users at every node/location based on bandwidth capacity of neighbouring links • Backward workload extrapolation (*): collects workload measurements from every node to the inner-code nodes • Inner-core workload extrapolation: extrapolates workload at every inner-core node (if needed) • Workload propagation (**): distributes the workload from inner-code nodes to every node in the network
Network-Routing-based		• All required by Hierarchy-based diffusion algorithm • A set of service (inner-core) nodes	• Routing path discovery: identifies (shortest) routing paths from client-clusters to the service nodes • User aggregation based on routing paths • Backward workload extrapolation (same as (*)) • Workload propagation (same as (**))

Experiments

Settings

- Network model
  - A small scale of the BT core network
    - 3 inner-core, 6 outer-core, 9 metro, and 27 T1 nodes
  - Distribution of nodes and assumptions of links' bandwidth capacity
    - Based on census population data of the city

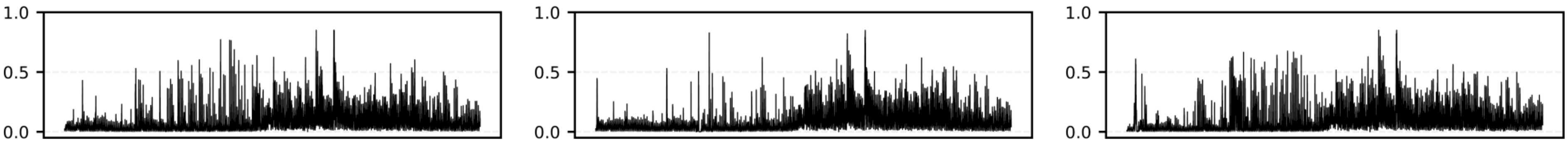
- Workload data [4]
  - From the production CDN system of BT
  - 3 datasets collected at 3 inner-core nodes



Network model of the city of Umeå, Sweden

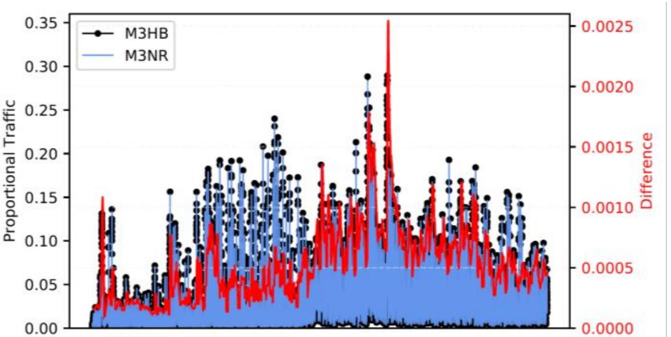
Experiments
Scenario 1 (1/2)

- Description
  - Measurements: at central nodes (I1, I2, I3)
  - Demonstration of basic features of the algorithms
    - Propagation of workload towards the edge of the network
  - Data is normalized; y-axis is named ’*Proportional Traffic*’
- Data traces

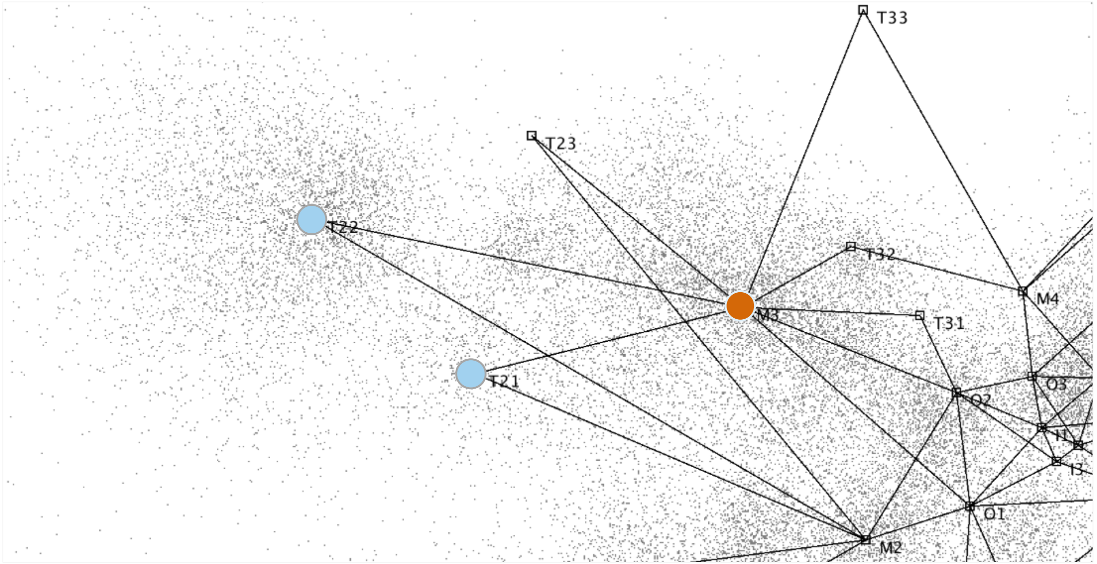


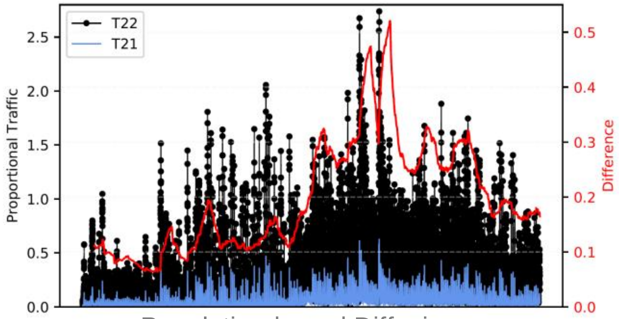
Original workload measurements associated to nodes I1, I2, and I3

Experiments
Scenario 1 (2/2)

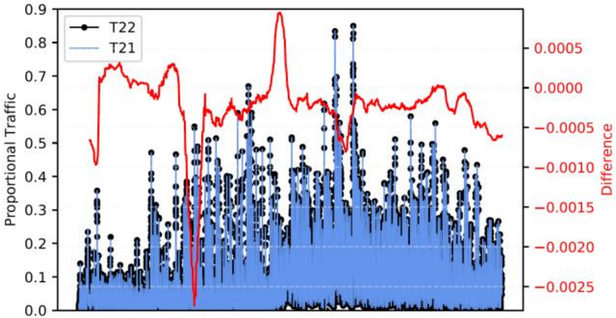


Hierarchy-based vs. Network-routing-based

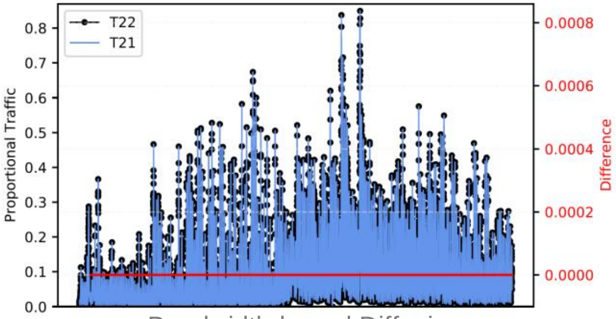




Population-based Diffusion



Location-based Diffusion



Bandwidth-based Diffusion

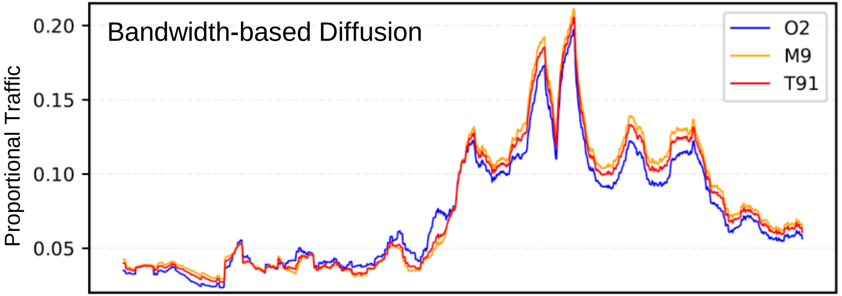
tieto

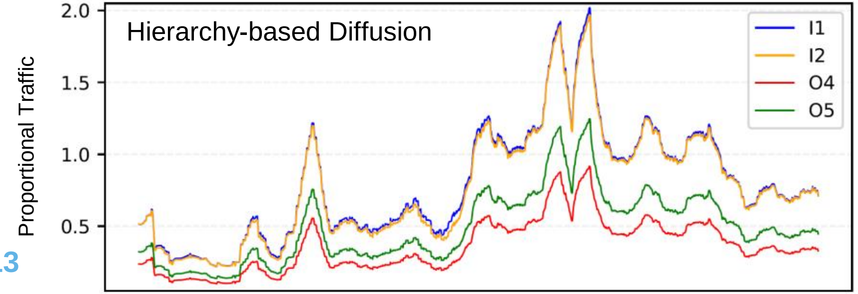
Experiments

Scenario 2

- Measurements: at random nodes (M2, M9, T62)
- Comprehensive verification

Workload patterns of nodes



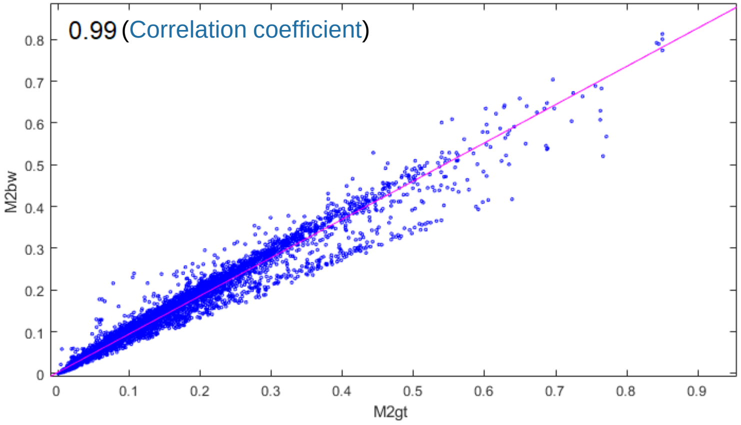


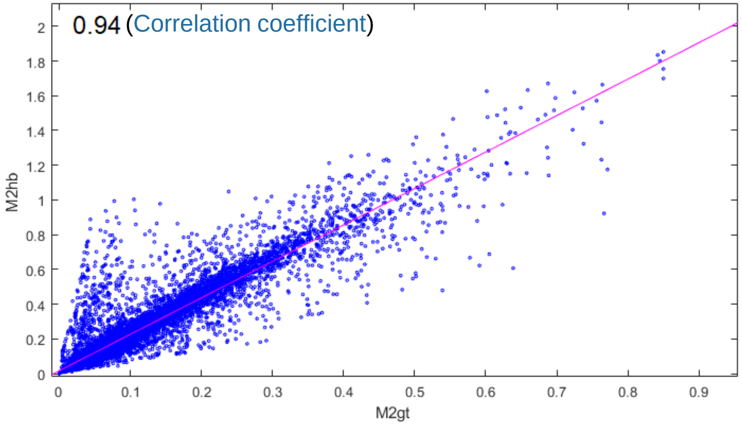
Result Validation

	M2	M2HB	M2BW	M9	M9HB	M9BW	T62	T62HB	T62BW
Entropy [5]	5.6252	6.7778	5.6368	5.7002	7.0710	5.6491	5.7588	6.8891	5.6464
Approximate Entropy [6]	0.6017	0.6344	0.6236	0.5972	0.6245	0.6202	0.6179	0.6257	0.6236

Entropy and approximate entropy measurements for the rediffused data of nodes M2, M9, and T62

(**BW**: bandwidth-based diffusion; **HB**: hierarchy-based diffusion)

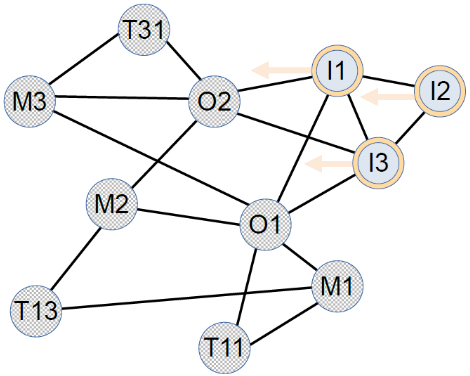




The distribution of the rediffused values and the original measurements for node M2

Discussion

- Main objectives
  - Workload generation to support large-scale distributed application profiling
  - Workload propagation modeling and/or application modeling
- Extension
  - Mitigate data privacy concerns in dissemination of data traces collected from sensitive data applications
    - E.g.: the scenario of BT CDN system (see the figure)
      - Core nodes I1, I2, I3: real measurements
      - Other nodes: generated data



tieto

Conclusions

- A formulation of the problem of workload generation for large-scale distributed applications/systems
- Five algorithms
  - Addressing the problem
  - Facilitating workload generation using workload propagation models
- A discussion on further application of the proposed diffusion algorithms

- Future work
  - To develop application models for telco service function chains and IoT applications
  - To develop or adapt the algorithms to the applications models: application profiling and data privacy
  - To standadize and abstract the models and algorithms to finalize a workload propagation modeling and workload generation framework

References

[1] T. Le Duc, R. Garcia Leiva, P. Casari, and P-O. Östberg. 2019. Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. ACM Computing Surveys 52, 5, Article 94 (August 2019), 39 pages.

[2] P-O. Östberg et al. 2017. Reliable Capacity Provisioning for Distributed Cloud/Edge/Fog Computing Applications. In Proc. European Conference on Networks and Communications (EuCNC). Oulu, Finland, 1–6.

[3] Kitz. 2009. BT 21CN – Network Topology & Technology. https://kitz.co.uk/adsl/21cn_network.htm. Accessed: February 19, 2020.

[4] M. Leznik et al. 2019. RECAP Artificial Data Traces. https://doi.org/10.5281/zenodo.3458559

[5] T. J. Ulrych and R. W. Clayton. 1976. Time Series Modelling and Maximum Entropy. Physics of the Earth and Planetary Interiors 12, 2 (1976), 188 – 200.

[6] S. Pincus. 1995. Approximate Entropy (ApEn) as a Complexity Measure. Chaos: An Interdisciplinary Journal of Nonlinear Science 5, 1 (1995), 110–117.

Thank you

Thang Le Duc, PhD
Postdoc Researcher
Department of Computing Science, Umeå University
Senior Researcher
Tieto Product Development Services, TietoEVRY Sweden AB
thang@cs.umu.se, thang.leduc@tieto.com