IBM

Towards data-driven NAND flash controller development

Roman Pletka, Nikolas Ioannou, Nikolaos Papandreou,
Radu Stoica, Saša Tomić, Haralampos Pozidis

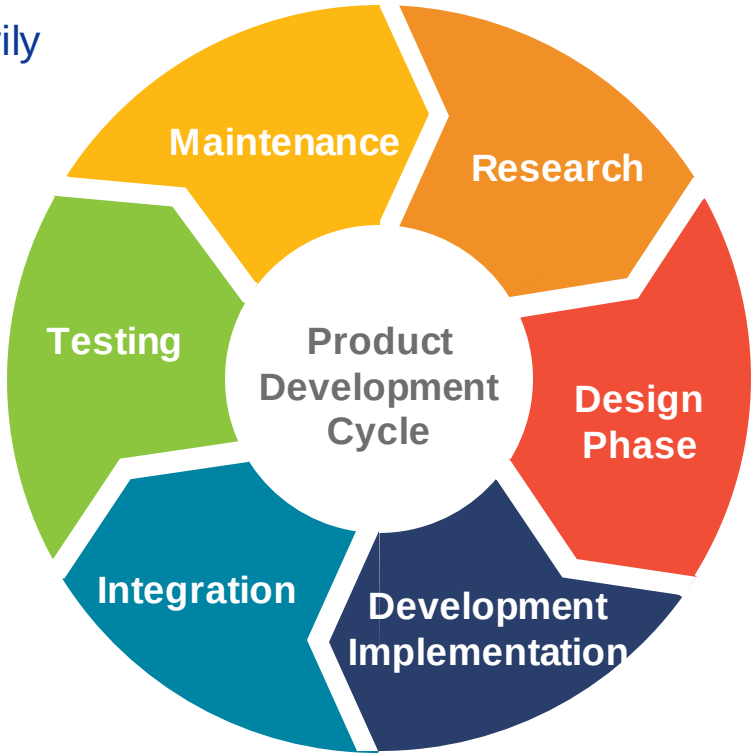
IBM Research – Zurich Research Laboratory

Outline

- Background
  - Controller development cycle
  - Challenges in current controller designs for QLC NAND Flash
- Data-driven controller development
  - How field data can deliver insights for future controller designs
- Deep dives:
  - Write amplification analysis
  - Write heat separation performance
- Conclusion

Controller development

- Speed of the controller development cycle is primarily dictated by the NAND Flash evolution:
  - New NAND flash generation every 12-18 month => 4 generations of 3D NAND in ~5 years
  - Availability of one generation only a few years or even less than a year
- Development of an ASIC controller can take more than 2 years. Design trade-offs:
  - Use of FPGAs instead of ASICs
  - Sub-optimal level shifting with increasing wear leads to higher tail latencies due to read retry or additional reads to gather soft information for LDPC
  - Less supported features (e.g., no multi-plane reads, write suspend, …)



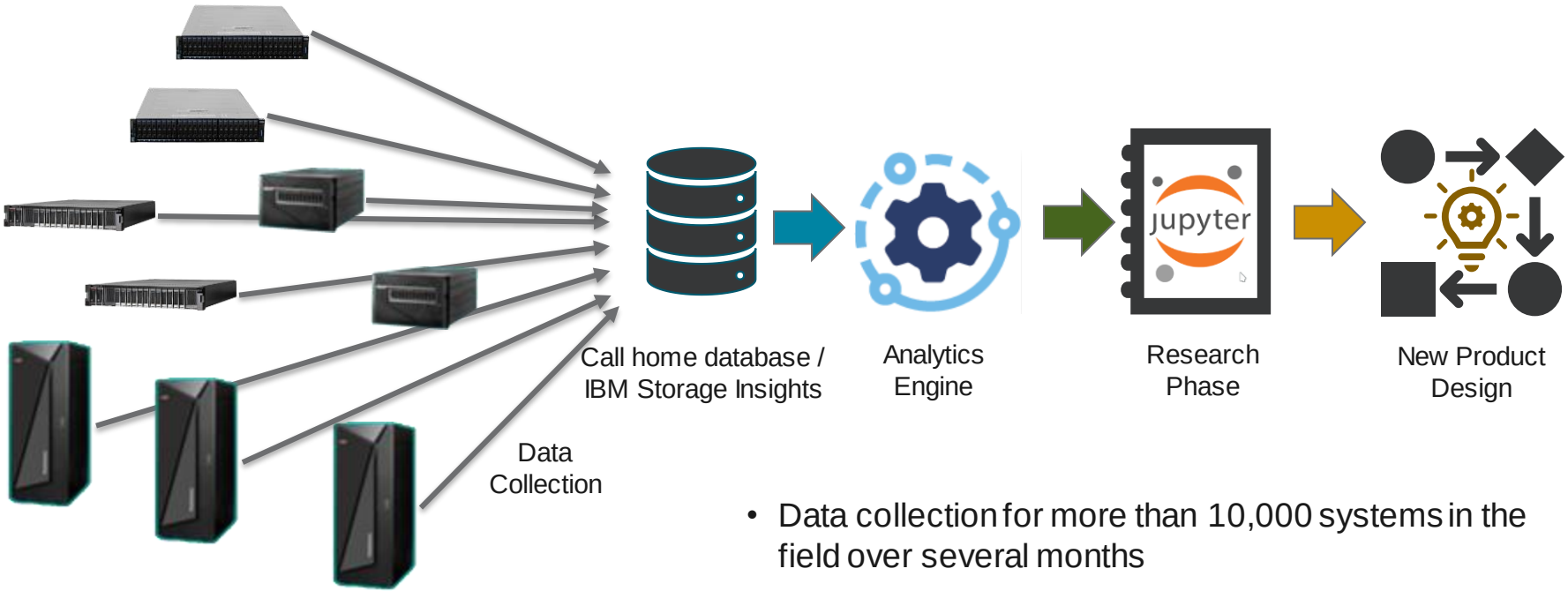
QLC controller challenges

- Latest 3D QLC NAND Flash faces many new challenges:
  - QLC has roughly 5-10x less endurance than TLC. Endurance is a challenge!
  - In QLC, of read, program, erase latencies increased by ~1.5 – 2x w.r.t TLC
  - Previous NAND Flash generation (e.g., 3D TLC) did not require SLC caching even for enterprise-level controllers
- This suggests that QLC should be preferably used for read intensive workloads. But read disturbs and retention also result in additional internal writes besides write amplification from garbage collection.
- What can we learn from previous controller generations when designing a QLC controller?

Data sources overview

- Call home data
  - Periodically collect system information describing the system hardware and critical configuration information
    - includes an excerpt from the event log, potential error events, contact information
  - Significantly improves the speed to resolve problems as it allows service personnel to contact customer and arrange service faster
    - No access to stored data
- IBM Storage Insights
  - Cloud-based storage management platform providing advanced analytics features
    - Managing complex storage infrastructures in a simple and integrated way
    - Enables efficient health, capacity and performance monitoring
  - Collection of storage infrastructure information and performance metrics
    - No access to stored data

Data-driven controller development



- Data collection for more than 10,000 systems in the field over several months

Outline

IBM

- Background
  - Controller development cycle
  - Challenges in current controller designs for QLC NAND Flash
- Data-driven controller development
  - How field data can deliver insights for future controller designs
- Deep dives:
  - **Write amplification analysis**
  - Write heat separation performance
- Conclusion

Write amplification factors

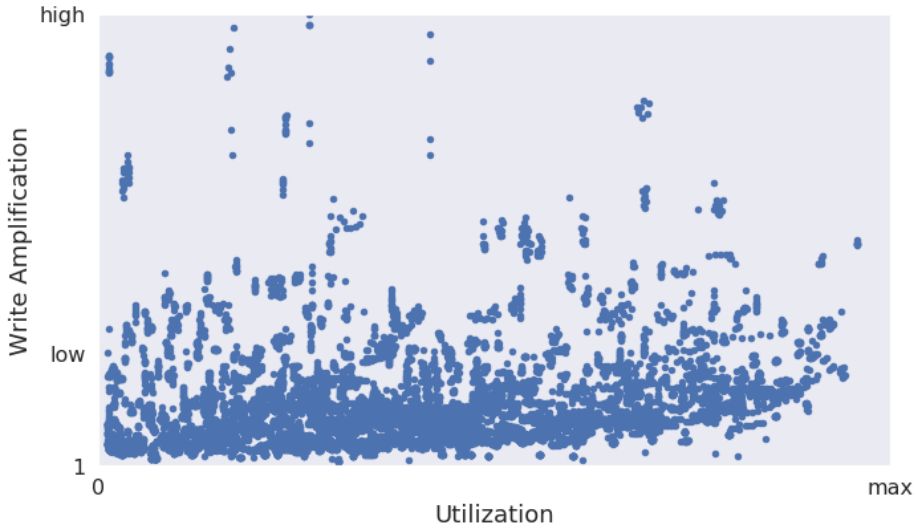
- Write amplification: $$WA = \frac{Total\ data\ written\ to\ Flash}{Total\ \ host\ writes}$$
- WA from garbage collection is a function of:
  - Overprovision and used capacity
  - Workload properties:
    - sequential, random, skewed writes
  - Write separation:
    - separating host and relocation writes
    - write heat separation (hot, warm, cold data)
  - Trim support
  - Relocations from retention and read disturb effects
- WA from retention and read disturbs is typically neglected

  ➥ How does retention and read disturb affect write amplification in real world?

Many analytical formulas, models, evaluations for WA exist: [Hu09, Agrawal10, Luojie11, Stoica13, Pletka18]

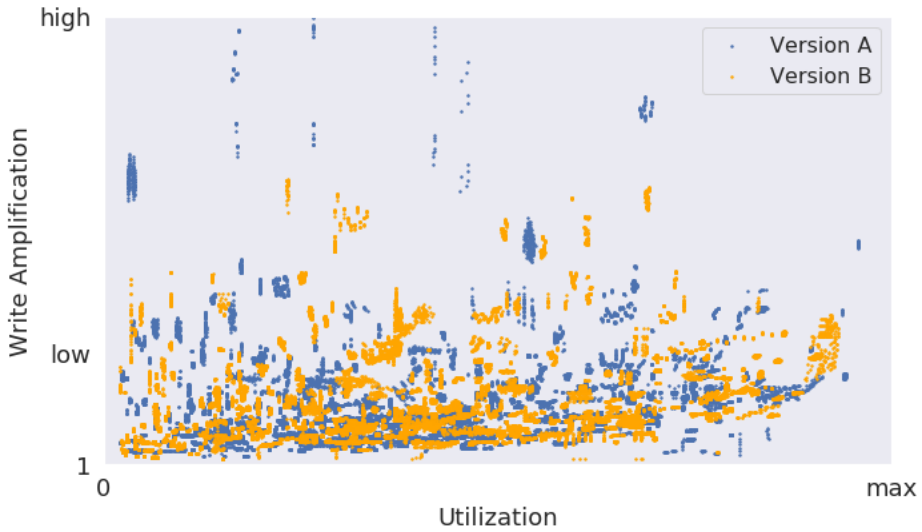
Measured Write Amplification

- Measured write amplification from more than 10k systems with 4-12 Flash Core Modules (FCM) per system
- Observation: additional internal writes cannot solely be contributed to host writes.
- There is a significant component from retention and read disturb effects which must be considered when designing a new controller.



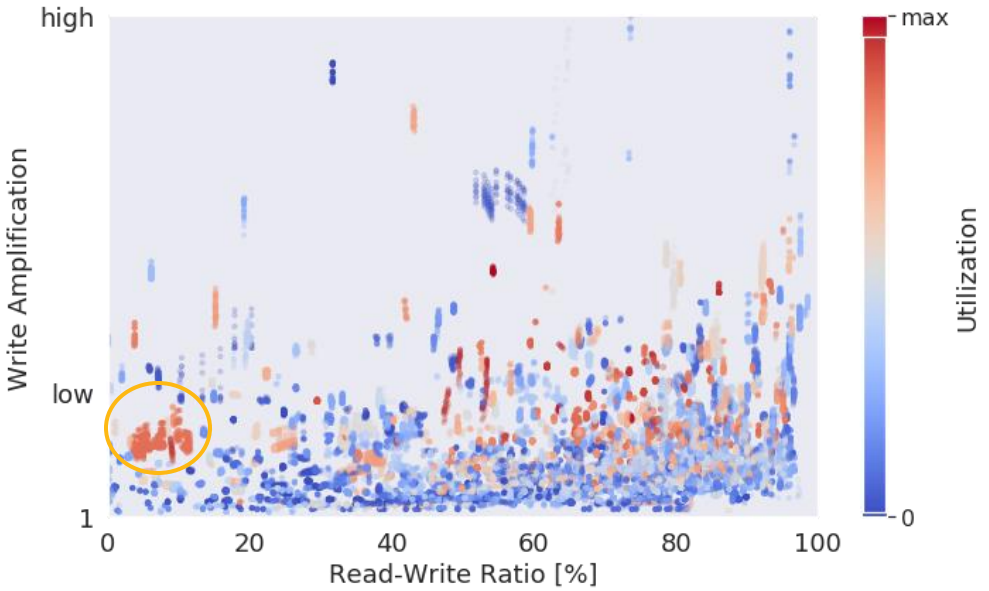
Measured Write Amplification

- Comparing the 2 most frequently used firmware versions
  - Version A: 37% of all systems
  - Version B: 32% of all systems
- Improvement in WA behavior of newer version B, but WA component from retention and read disturb effects are still dominating in most systems.



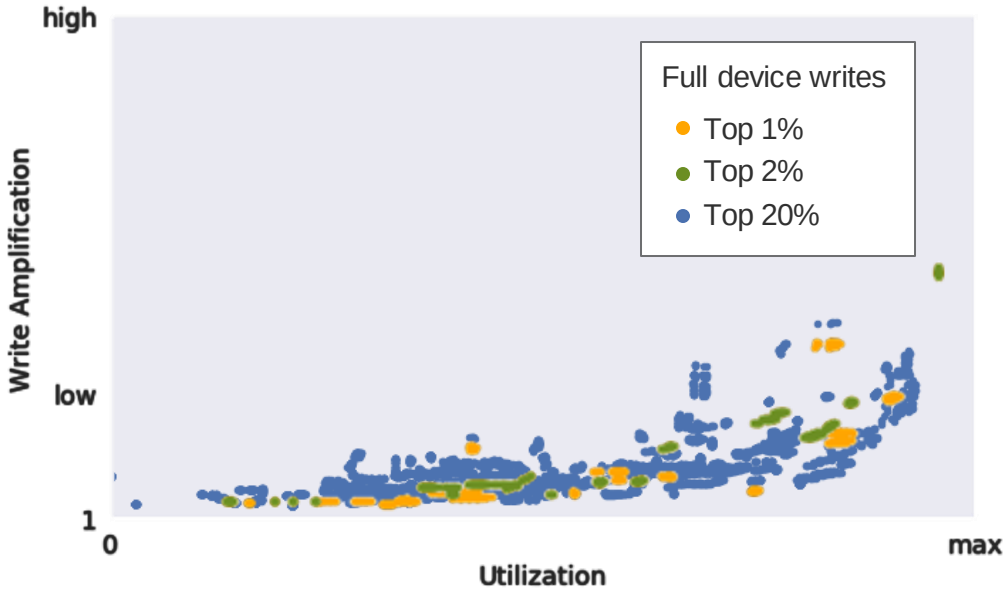
WA and host read-write ratio

- Higher fraction of writes tend to have more WA; however, this is not generally the case:
  - Clusters indicate systems with similar workload properties:
    - Example: systems with low read ratio and high utilization but low WA



WA from host writes

- Comparison of systems with top 1 / 2 / 20% highest amount of full physical device writes.
- With increasing full physical device writes, WA is dominated by garbage collection.
- Even though WA measured in most other systems is significantly higher, their total number of writes is still low (i.e., low impact on endurance).



Outline

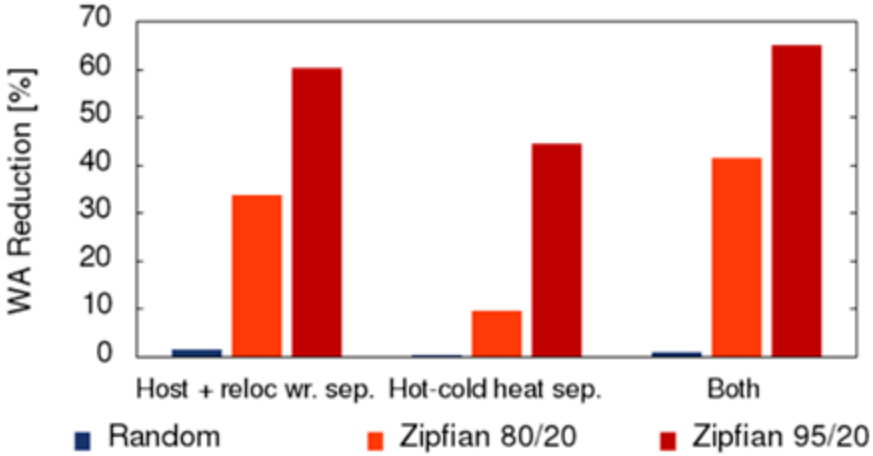
IBM

- Background
  - Controller development cycle
  - Challenges in current controller designs for QLC NAND Flash
- Data-driven controller development
  - How field data can deliver insights for future controller designs
- Deep dives:
  - Write amplification analysis
  - **Write heat separation performance**
- Conclusion

Write heat separation

Write heat separation (Wr-HS):

- Separation in 2 dimensions:
  - Host and relocation write separation
  - Separation of all writes according to their update frequency
- Tracking heat information on LBA or LBA range granularity
- Significant WA reduction reported using synthetic workloads and traces
  - Host and relocation write separation is more important than simple hot-cold separation
  - Both schemes together further reduce WA

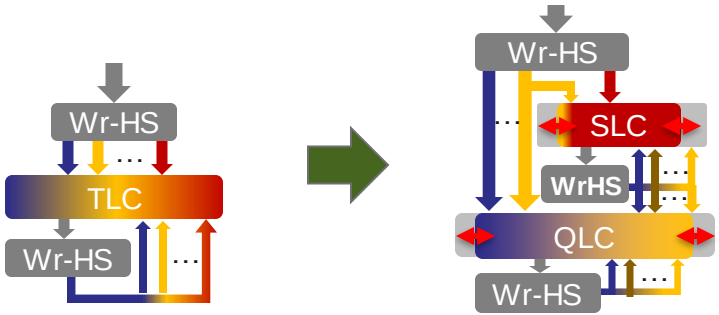


Source: [Pletka18]

Write heat separation

Towards a QLC controller design:

- The introduction of two dynamically sized pools adds significant complexity in a controller design

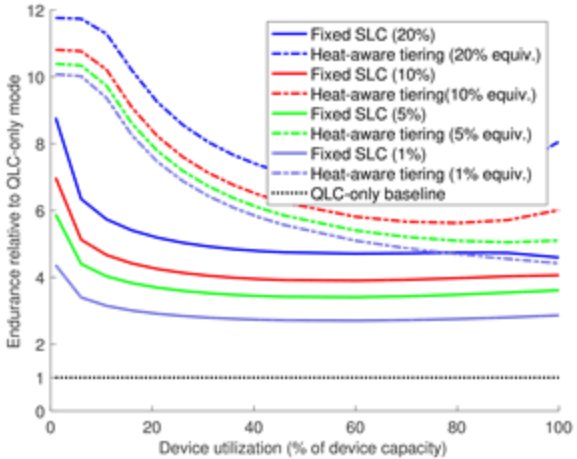


Evaluation of controller design alternatives:

- Modeling a hybrid SLC-QLC controller
- Validate model with data from real systems

Results from SLC-QLC controller model:

- Example: Zipfian 95/20 workload shows high endurance gains with optimal data placement even under high utilization

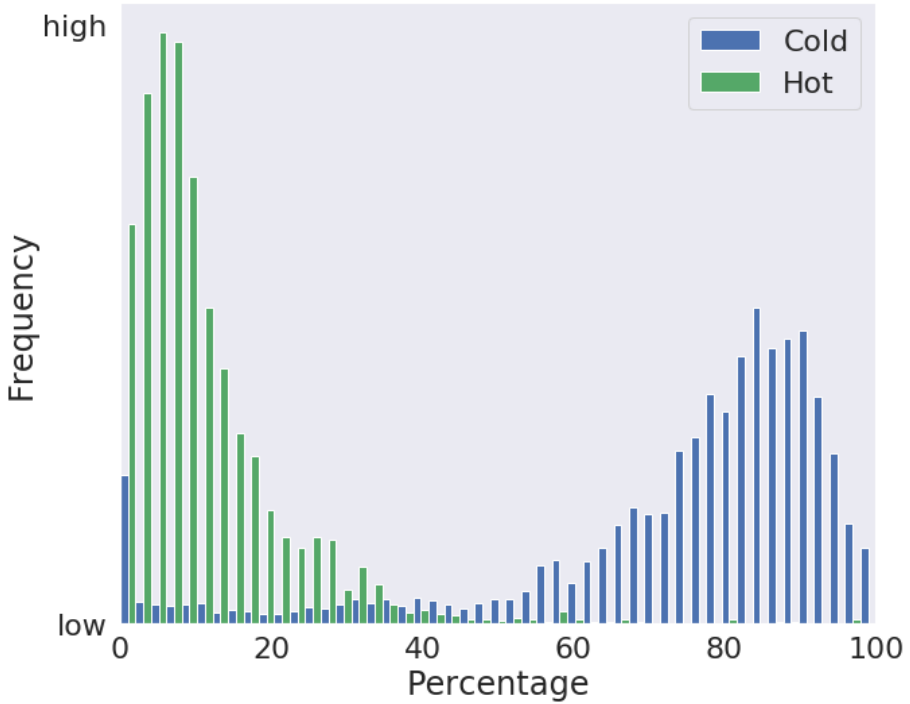


- See more detailed results in [Stoica19]

Measured write heat

What is the write skew in real-world workloads?

- Collecting heat information from all systems in the field using n heat streams: hot=1, …, cold=n
- Presenting only hot and cold streams. Observations:
  - On average, systems have 70% cold and 10% hot data.



Conclusion

IBM

- Field data is key to understand dynamics of real controllers at large scale
  - new features and enhancement are being delivered faster
- Data-driven controller development
  - enables a better understanding of key metrics for a new controller design
  - helps to bridge the gap between the controller development and NAND flash generation cycles
- Data-driven controller development must meet global data compliance standards:
  - EU-US Privacy Shield and Swiss-US Privacy Shield Framework
  - ISO 27001

References

- [Hu09] X.-Y. Hu, E. Eleftheriou, R. Haas, I. Iliadis, and R. Pletka, "Write amplification analysis in flash-based solid state drives", Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, 2009
- [Agrawal10] R. Agarwal, M. Marrow, "A closed-form expression for Write Amplification in NAND Flash", Proceedings of IEEE Globecom Workshop on Application of Communication Theory to Emerging Memory Technologies, 2010
- [Luojie11] X. Luojie, B. M. Kurkoski, "An Improved Analytic Expression for Write Amplification in NAND Flash" International Conference on Computing, Networking and Communications (ICNC), 2012
- [Desnoyers12] P. Desnoyers, "Analytic modeling of SSD write performance", Proceedings of the 5th Annual International Systems and Storage Conference, 2012
- [Stoica13] R. Stoica and A. Ailamaki, "Improving flash write performance by using update frequency", Proceedings of the VLDB Endowment, vol. 6, no. 9, 2013
- [Pletka18] R. Pletka, I. Koltsidas, N. Ioannou, S. Tomić, N. Papandreou, T. Parnell, H. Pozidis, A. Fry, T. Fisher, "Management of next-generation NAND flash to achieve enterprise-level endurance and latency targets", ACM Trans. Storage, vol. 14, no. 4, Dec. 2018. http://doi.acm.org/10.1145/3241060
- [Stoica19] R. Stoica, R. Pletka, N. Ioannou, N. Papandreou, S. Tomic, H. Pozidis, "Understanding the design trade-offs of hybrid flash controllers", To appear in proceedings of the 19th IEEE Int. Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems MASCOTS, Oct. 2019

Thank You !