Session 3 - Summary
Topic: AI Applications

Rapporteur:
Jiangshan Yu
The University of Sydney

Acknowledge:
Behrooz Sangchoolie
RISE Research Institutes of Sweden, Sweden

THE UNIVERSITY OF SYDNEY

Talk 1

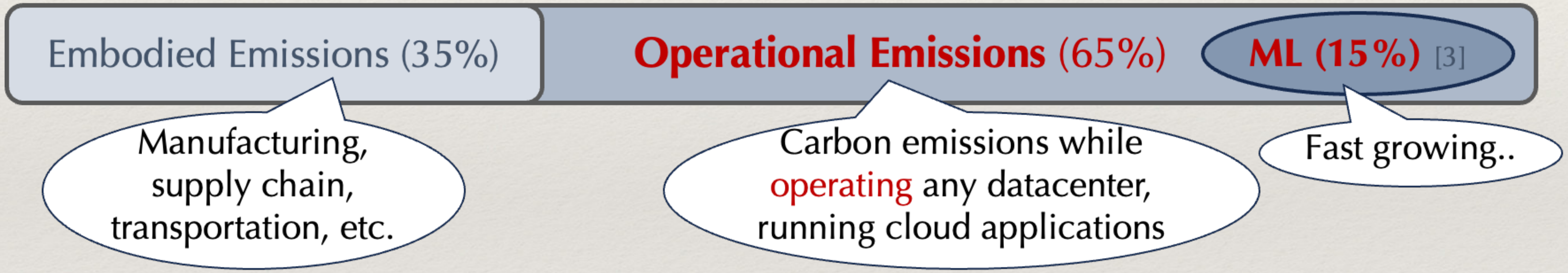
When Green Computing Meets Performance and Resilience SLOs

Ravishankar K. Iyer
UIUC, USA

THE UNIVERSITY OF SYDNEY

Motivation - Carbon emission

- Massive cloud systems consume a lot of energy
- Cloud for ML is power hungry
  - United nation: Net Zero by 2050: the world's most urgent mission



  - Google: 60% carbon footprint goes to model serving [2021]

THE UNIVERSITY OF SYDNEY

Why relevant to IFIP WG 10.4?

- ML fails when the input is different from the distribution for training the model. This then cost even more power.
- Resilience to **ML failures** and **classic faults** is not cheap
  - Fault management: 40-60% energy consumption overhead
- Carbon footprint optimization can lead to service level objectives (SLOs) violations.
  - Availability — always deliver whenever needed — is a legal requirement in Australia. Requires fault management for green computing, due to this availability requirements. Fault management for ML in the cloud would therefore be important.

THE UNIVERSITY OF SYDNEY

Key points

- Can we rely on batteries?
  - All green energy (e.g., solar, wind) has fossil fuel consumption
  - Cost of resilience: Requires substantial cloud management efforts
  - Power storage cost can be very high -- trillions of dollars
- Sustainability Challenge:
  - Requires significant new interdisciplinary research from SysML & resilience communities

THE UNIVERSITY OF SYDNEY

Key points (cont.)

- Two goals [or, rather, a trade-off]:
  - Sustained energy and sustained performance (including resilience)

- Key question: How to address large system+ML resilience management?

- [DSN-Distupt’24] Introducing *μ-serve model serving*, leveraging game theory, show how to reduce the energy consumption from top to bottom.
  - Achieves 1.2-2.6x higher power saving.

When Green Computing Meets Performance and Resilience SLOs. Haoran Qiu, Weichao Mao, Chen Wang, Saurabh Jha, Hubertus Franke, Chandra Narayanaswami, Zbigniew T. Kalbarczyk, Tamer Başar, Ravishankar K. Iyer. E Energy 2024 Singapore June ‘24.

THE UNIVERSITY OF SYDNEY

Q&As:

- Given the level of availability, reliability, how much energy could be minimally spent?
  - No research available;
  - The base-level numbers from the vendors are also not available.
- Would saving cost encourages more frequent usages? Energy consumption and carbon emission is different
- Other approaches (e.g. Life-cycle analysis) could be taken into account to solve optimization problem.

THE UNIVERSITY OF SYDNEY

Talk 2

Blockchain Room of Requirements (BR^2): an LLM-Enhanced Simulator for Blockchain Protocols

Cong Wang
City University of Hong Kong, China

THE UNIVERSITY OF SYDNEY

Motivation - blockchain education

- Challenges for students to play around:
  - Blockchain evolves very fast;
  - Many different attacks
- Hardhat: An Ethereum development environment
  - For creating/test/replaying smart contracts
  - Time consuming to create even one single configuration

Key points

- **Goals** (for teaching assistants):
  - Streamlined Custom Configuration
  - Intended Transactions
- **Challenge**: general-purpose LLMs fail in domain-specific tasks
- **Key lessens:**
  - External knowledge (EK) is critical in the optimisation
  - Leveraging standard Retrieval-Augmented Generation (RAG)
    - Embedding (1) Hardhat configuration template and (2) contract source code as EK, to support query and search, resp.
    - Result: Simple request in human language is enough e.g. update to change the configuration in getting more values

THE UNIVERSITY OF SYDNEY

Future challenges/work:

- Build the benchmarking dataset;
- Evaluation (w/o ground truth);
    - Bias: generated from scratch by human v.s. assessing a given LLM output
- Optimise RAG pipeline

THE UNIVERSITY OF SYDNEY