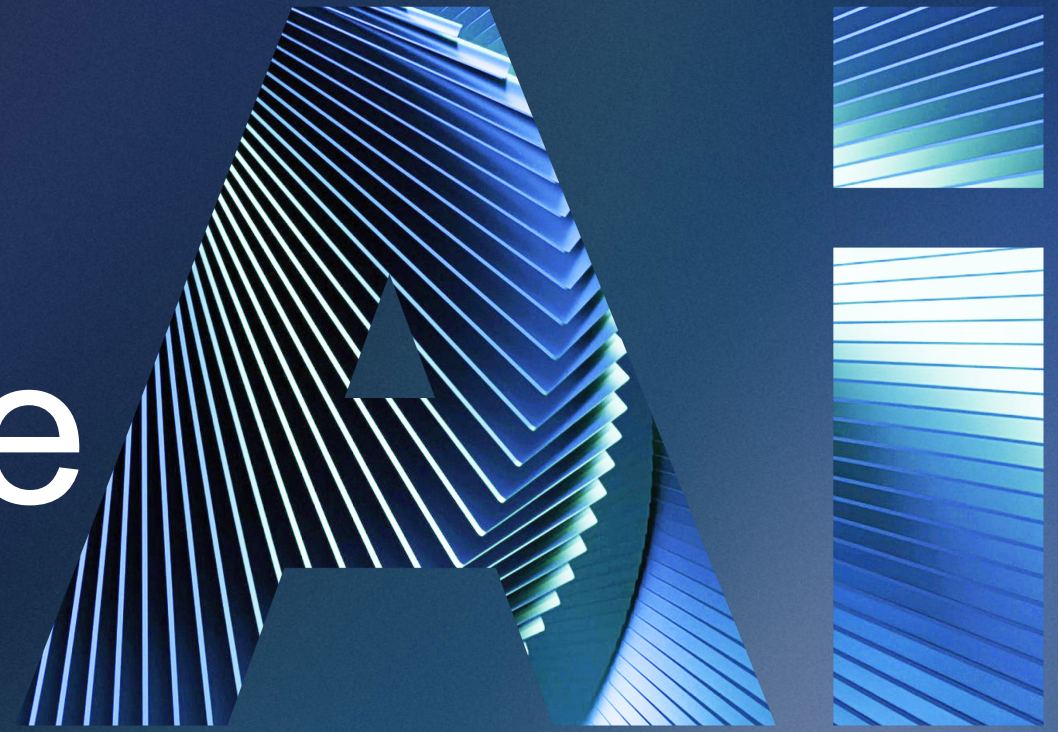
LLMs on Edge

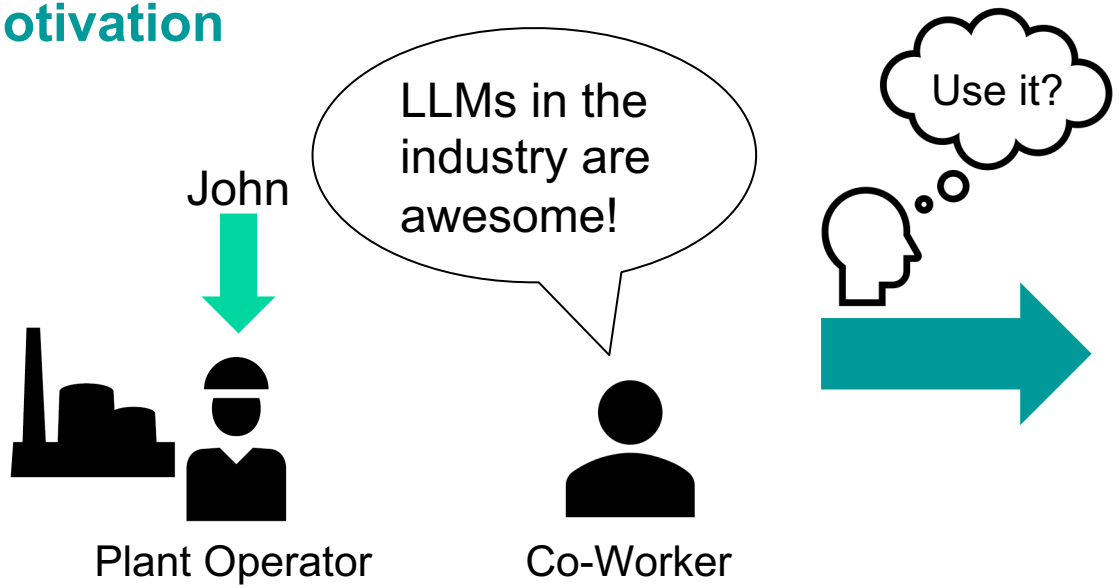


Network Traffic Characteristics of Distributed Inference under the Loupe

Philippe Buschmann, Arne Bröring, Georg Carle, Andreas Blenk

SIEMENS

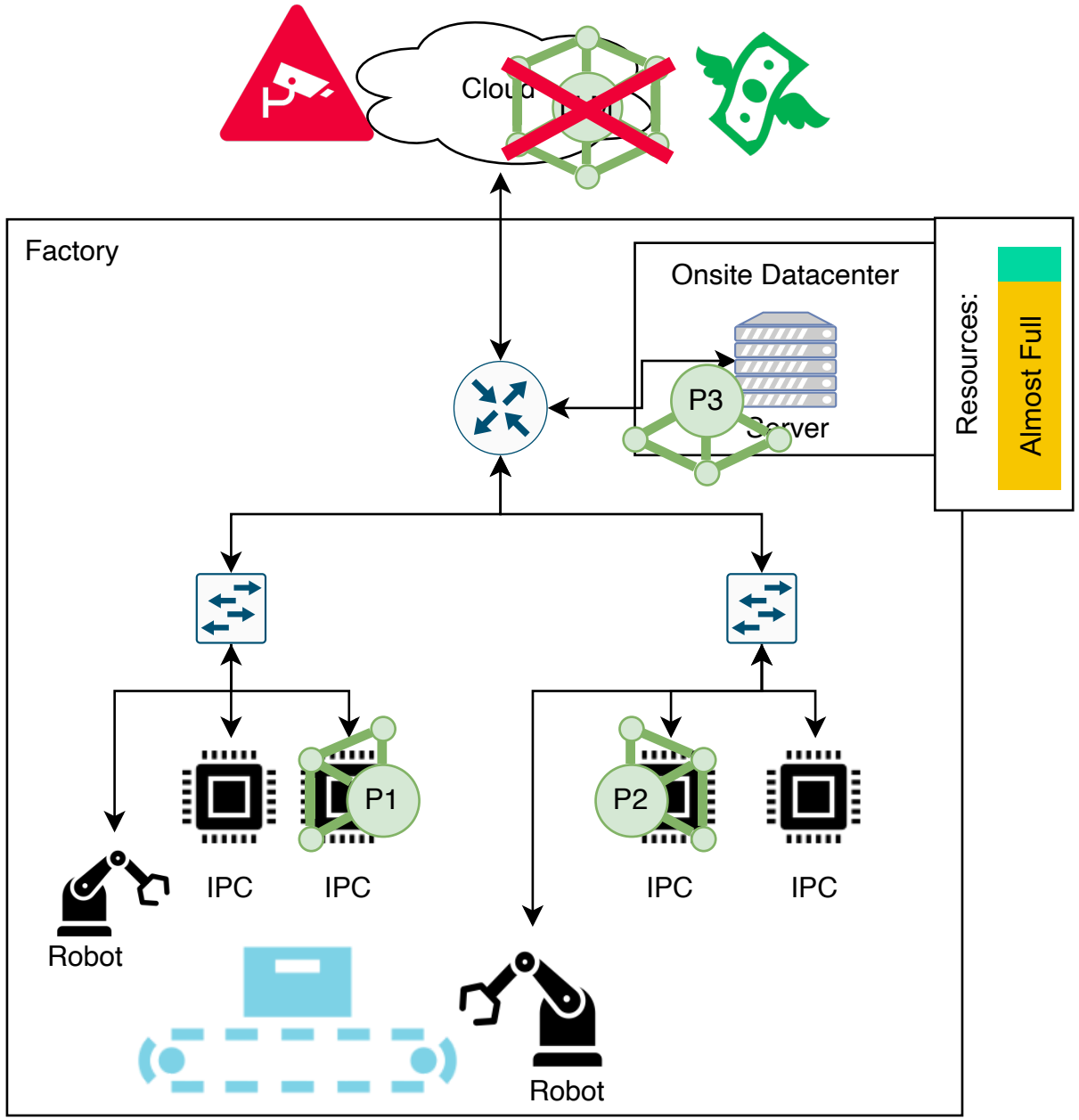
Motivation



Requirements / Challenges:

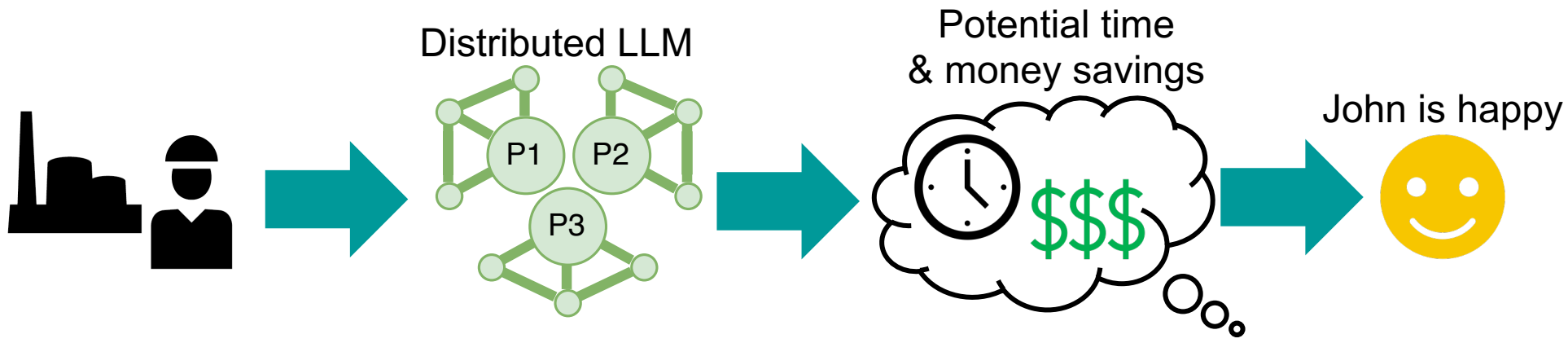
1. Privacy & Costs – LLMs on-premise only
2. Resource Constraints – On-premise datacenter not sufficient

➔ Distributed LLMs → Split up the inference of LLMs

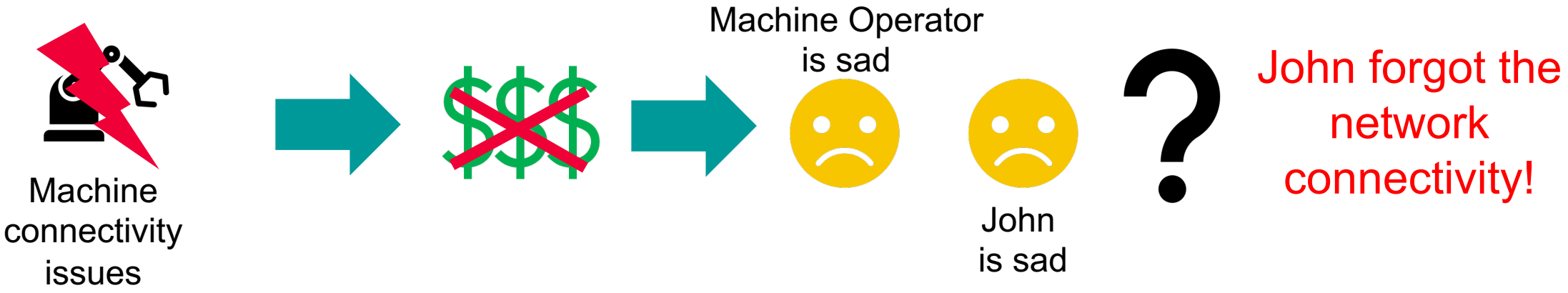


SIEMENS

Motivation

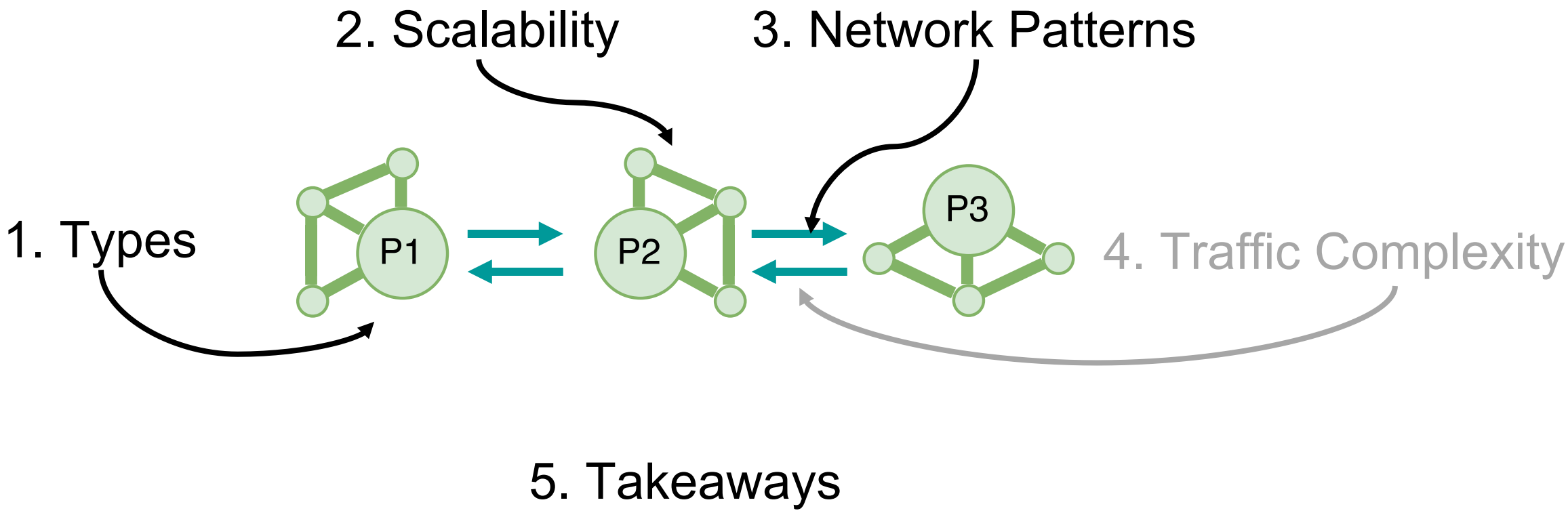


After some time:



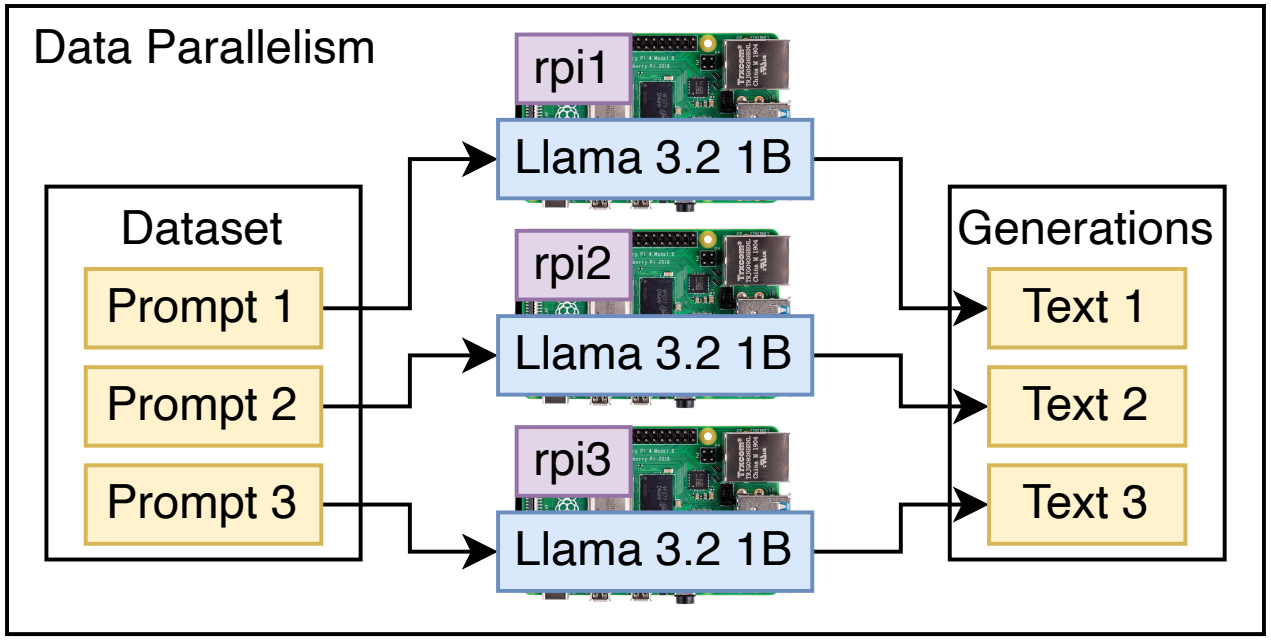
SIEMENS

Distributed Large Language Models



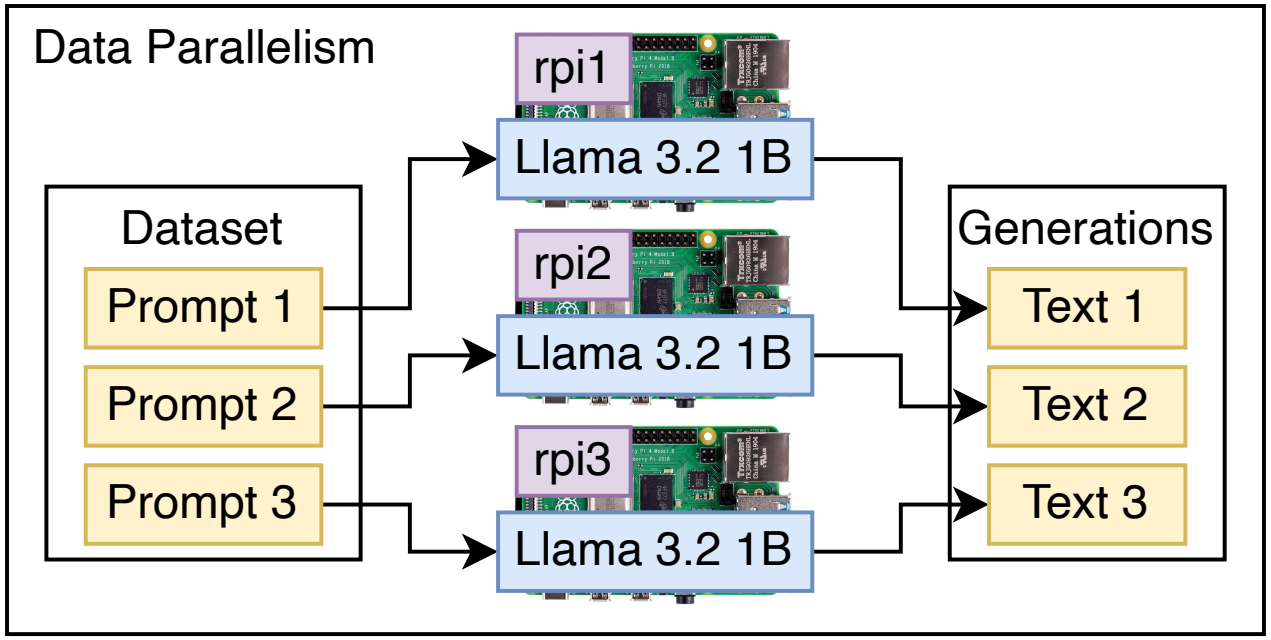
SIEMENS

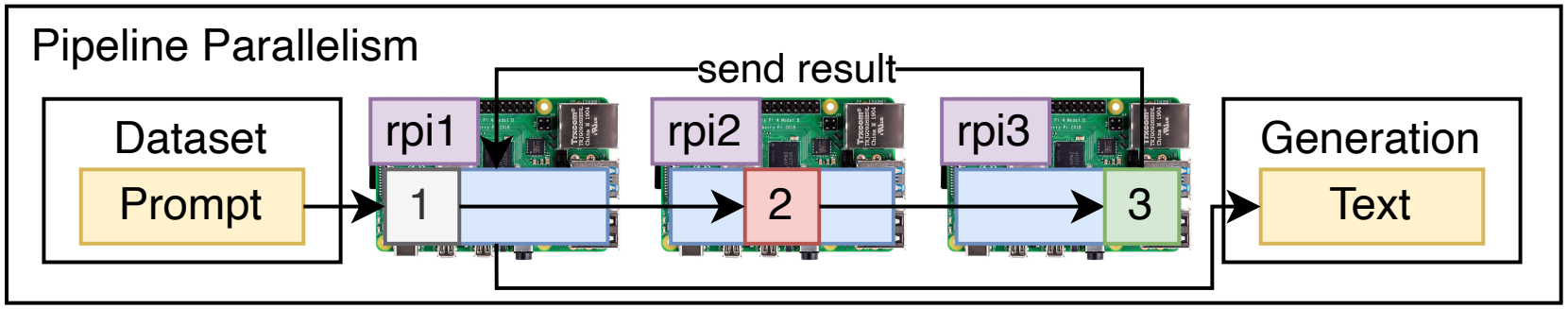
Types of Distributed Large Language Models



SIEMENS

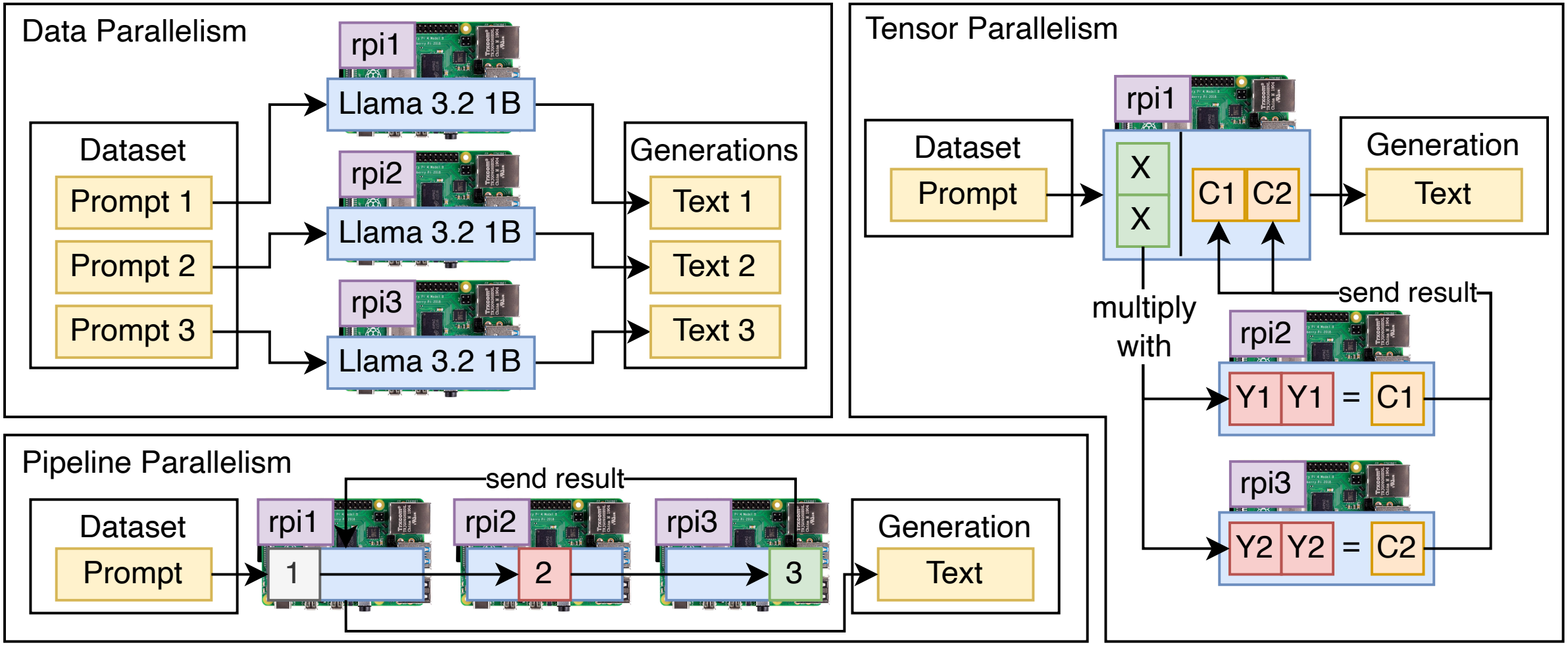
Types of Distributed Large Language Models





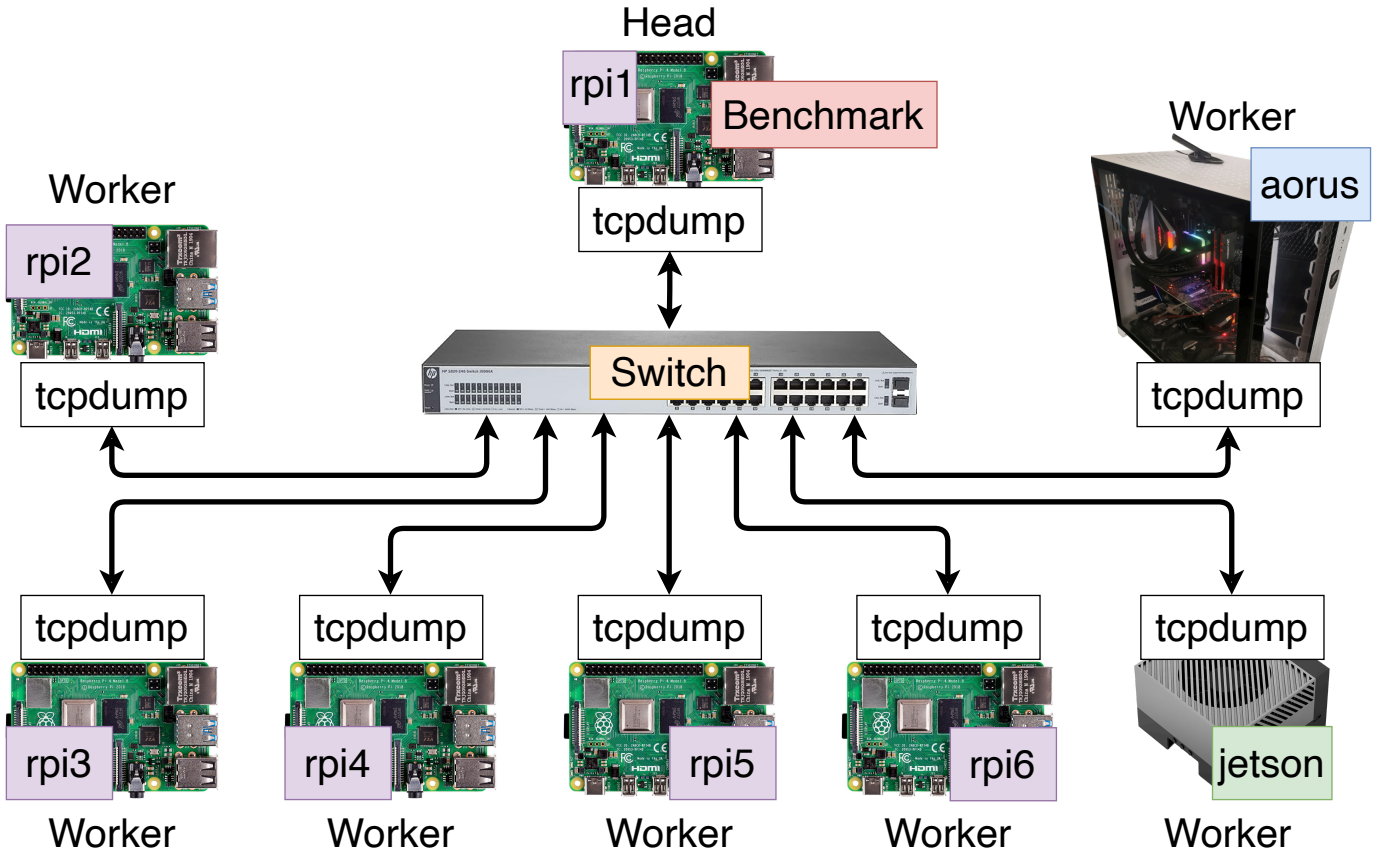
SIEMENS

Types of Distributed Large Language Models



SIEMENS

Testbed for Distributed LLMs



Pipeline Parallel Framework

`llama.cpp`

Tensor Parallel Framework

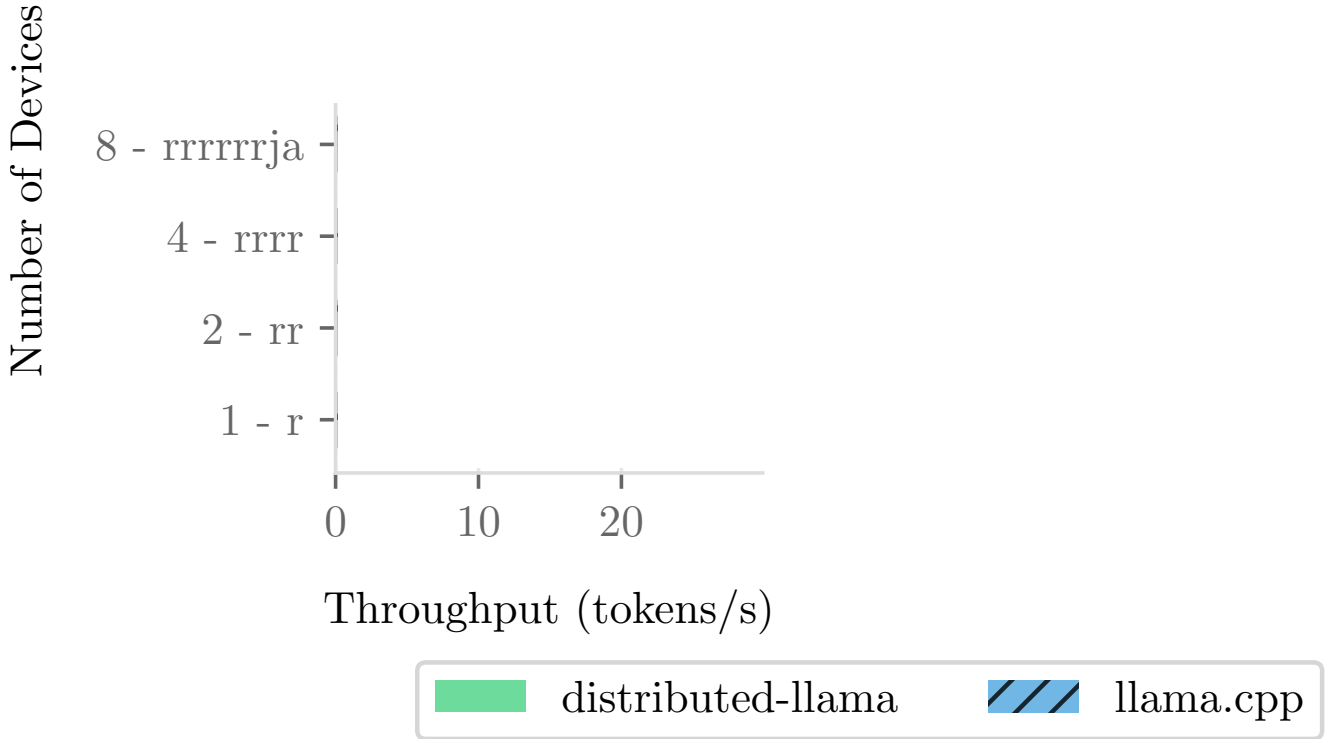
`distributed-llama`

SIEMENS

Results – Scalability

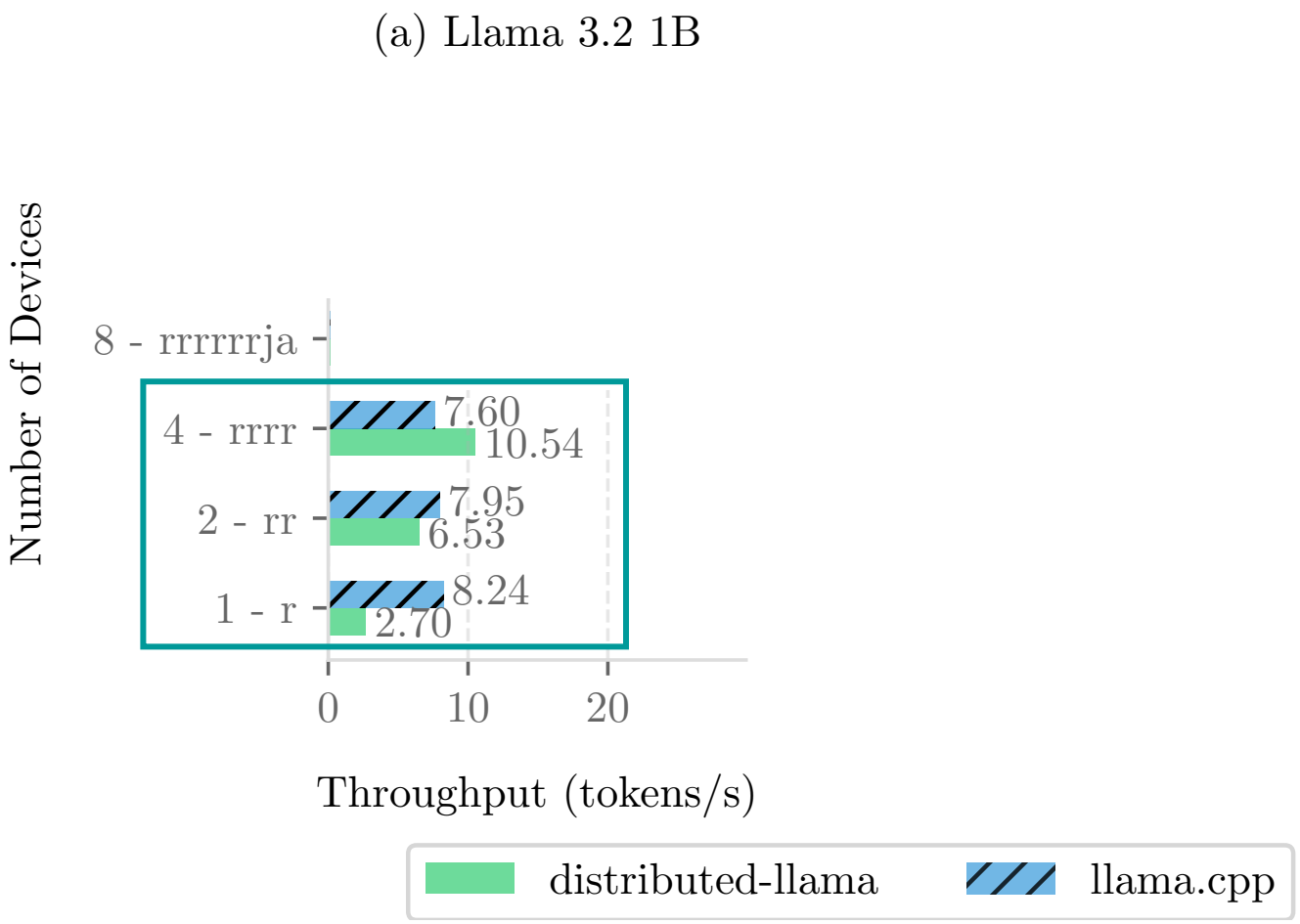
(a) Llama 3.2 1B

Results:



SIEMENS

Results – Scalability

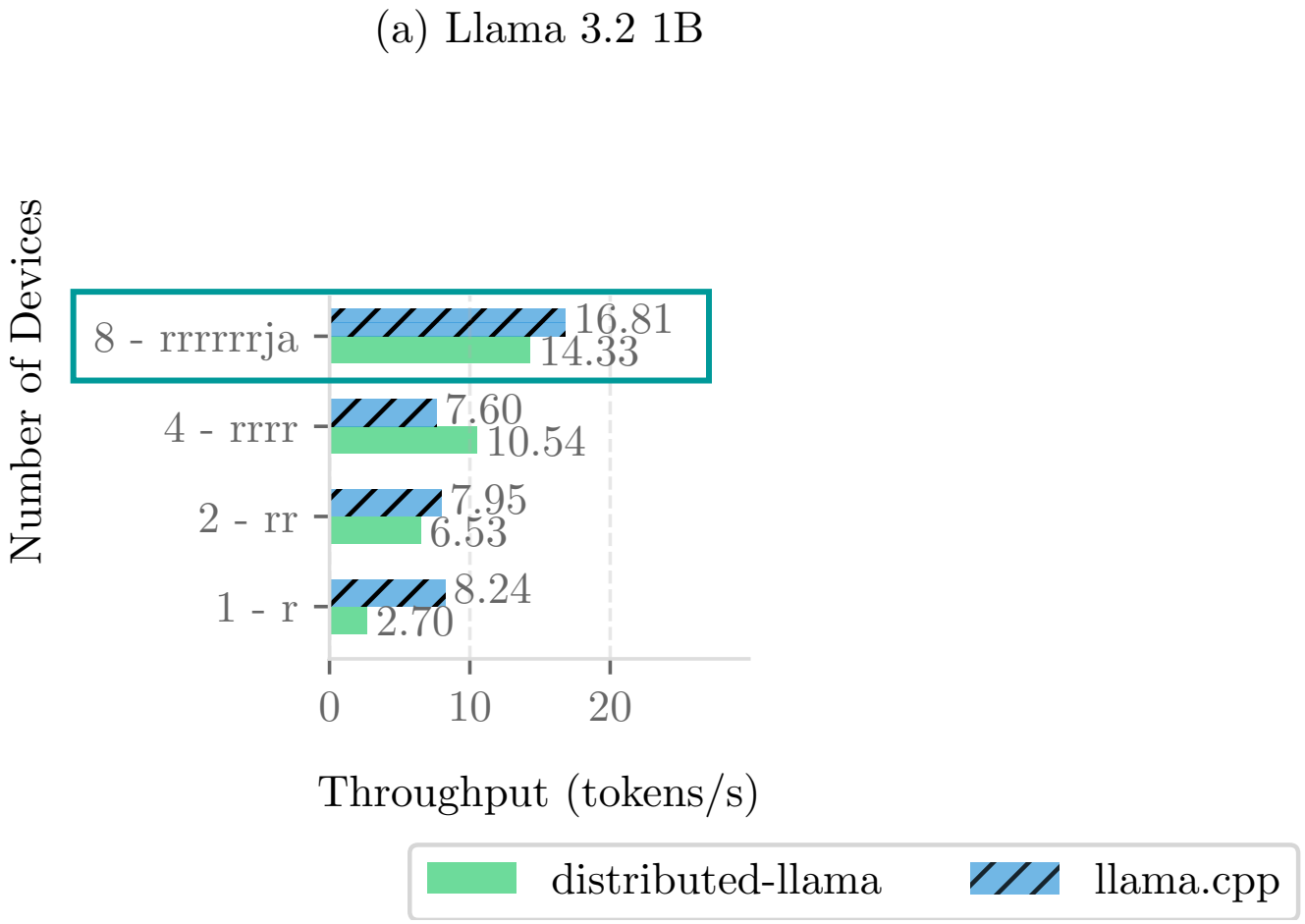


Results:

- `llama.cpp`:
  - More devices → Same throughput
- `distributed-llama`:
  - More devices → Higher throughput

SIEMENS

Results – Scalability

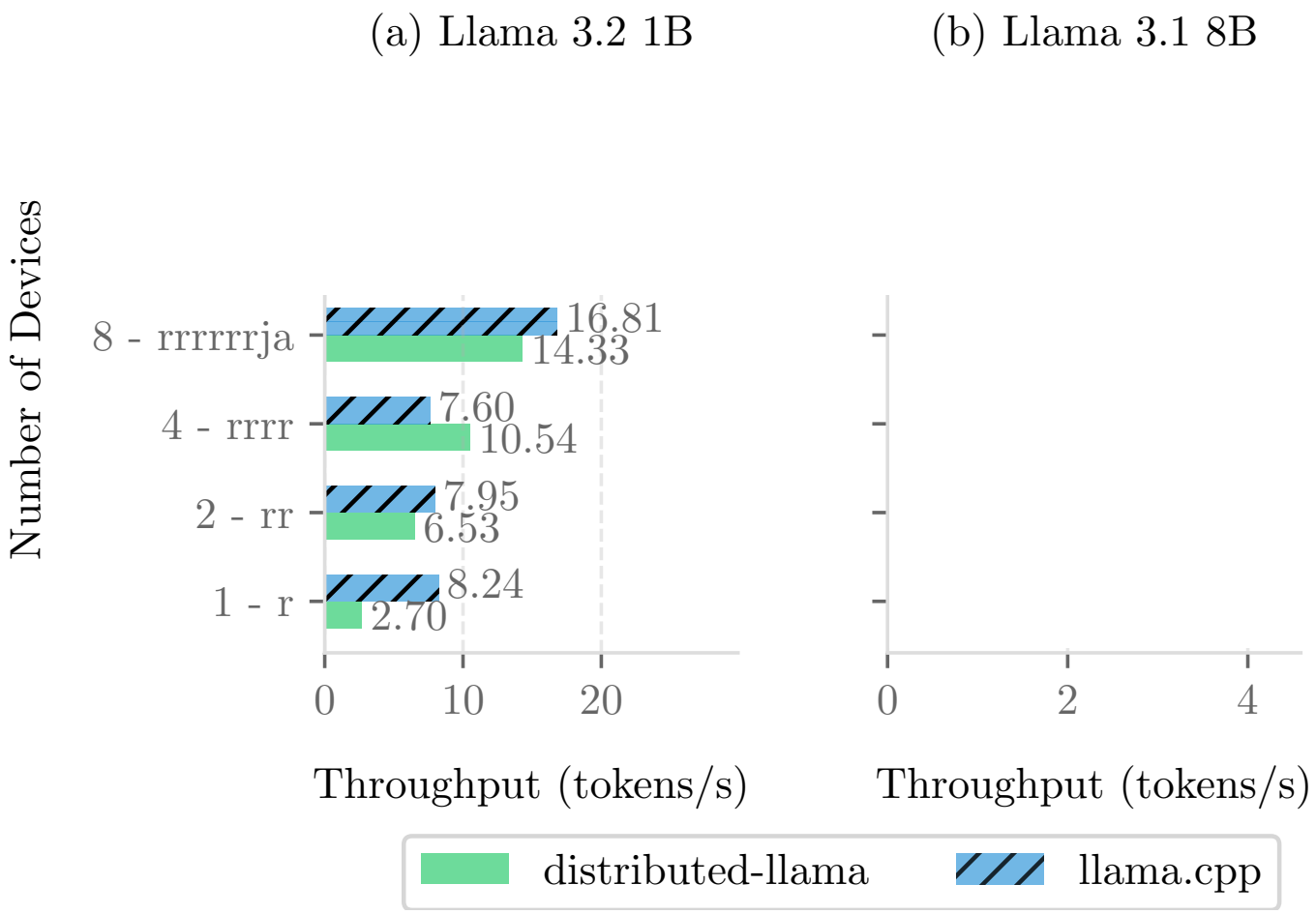


Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
- `distributed-llama`:
  - More devices → Higher throughput

SIEMENS

Results – Scalability

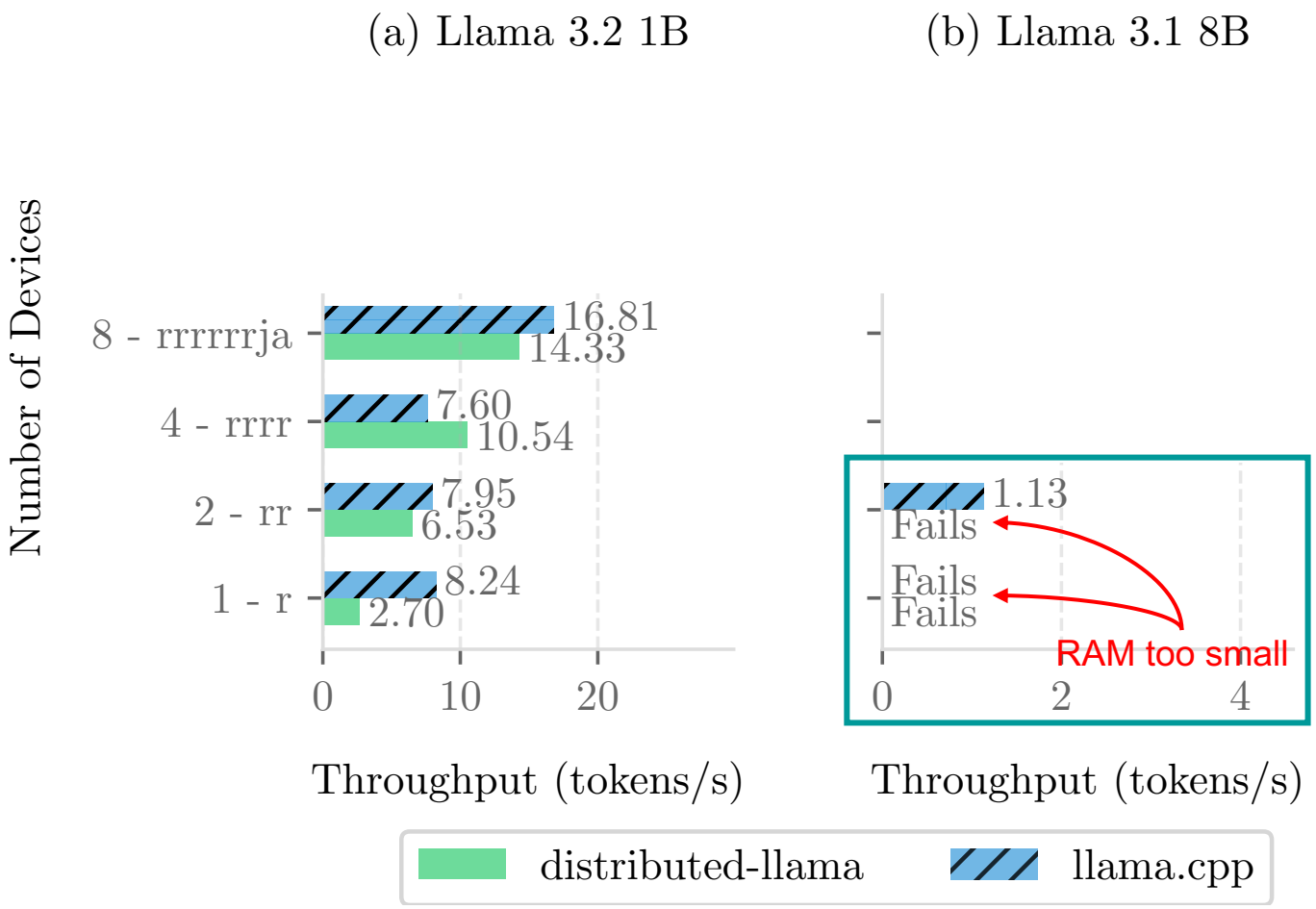


Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
- `distributed-llama`:
  - More devices → Higher throughput

SIEMENS

Results – Scalability



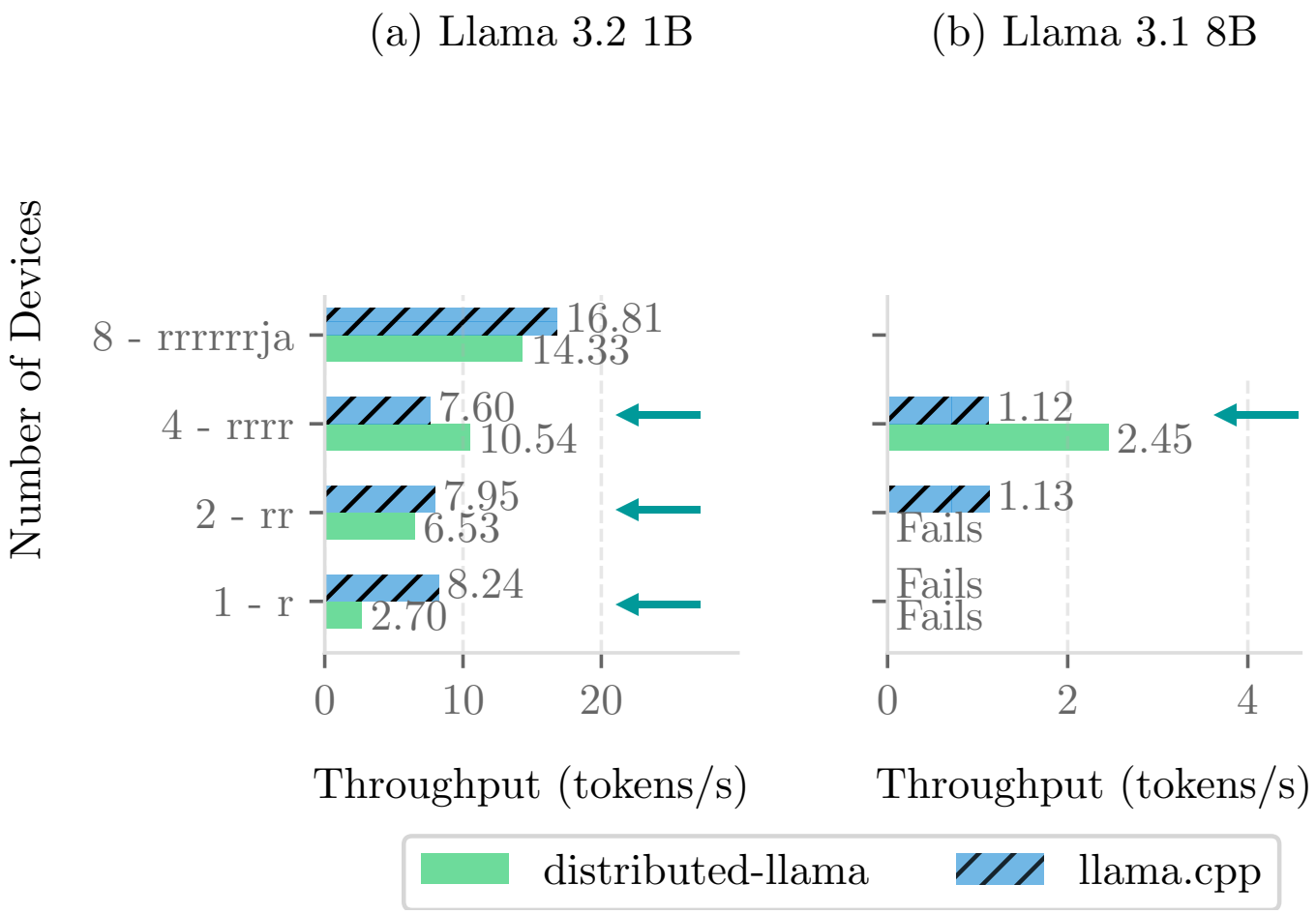
Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
  - More efficient RAM usage

- `distributed-llama`:
  - More devices → Higher throughput

SIEMENS

Results – Scalability



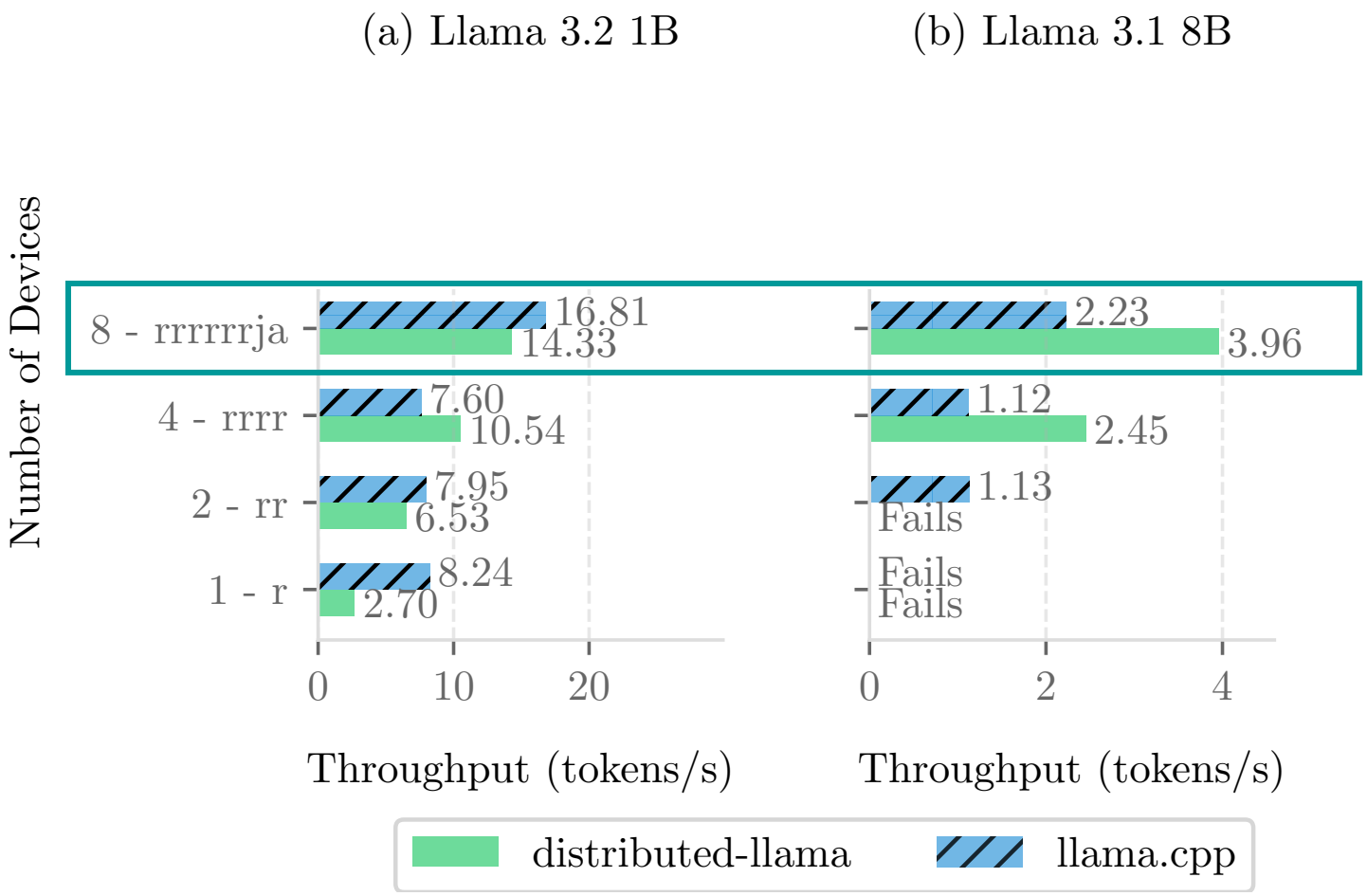
Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
  - More efficient RAM usage

- `distributed-llama`:
  - More devices → Higher throughput

SIEMENS

Results – Scalability

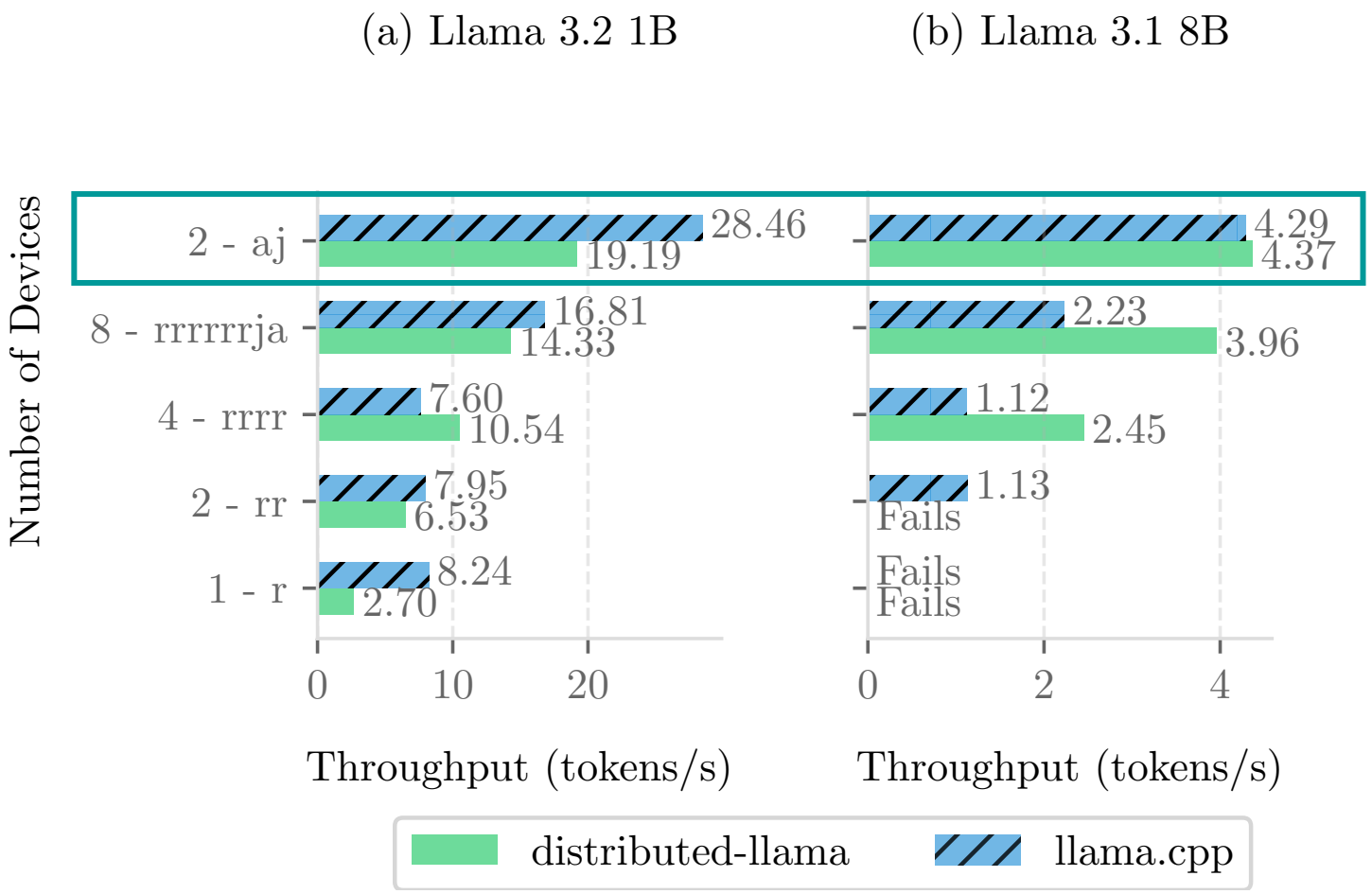


Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
  - More efficient RAM usage
  - Lower throughput for larger models
- `distributed-llama`:
  - More devices → Higher throughput
  - Higher throughput for larger models

SIEMENS

Results – Scalability

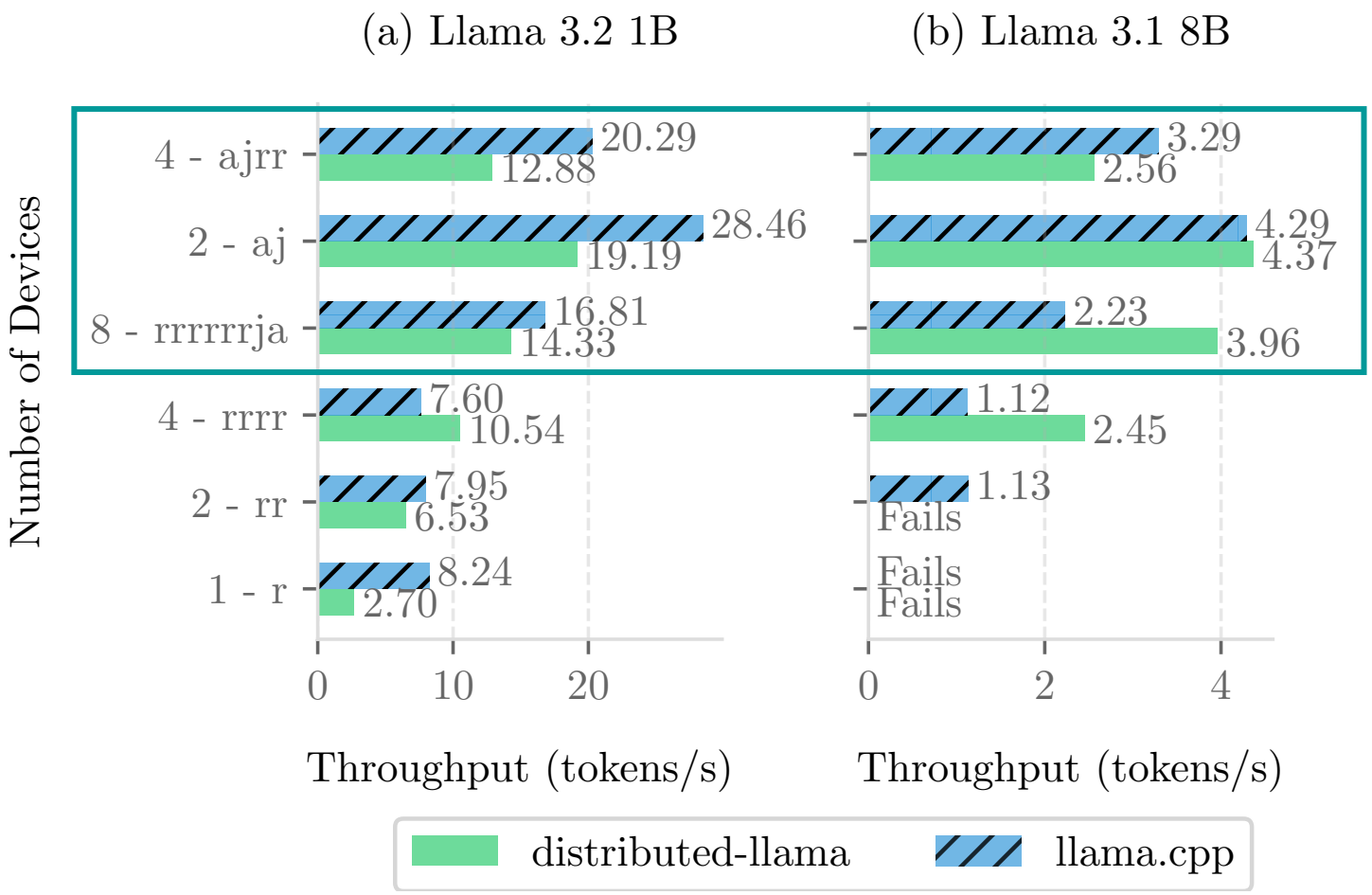


Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
  - More efficient RAM usage
  - Lower throughput for larger models
- `distributed-llama`:
  - More devices → Higher throughput
  - Higher throughput for larger models
- Both:
  - Better devices → Higher throughput

SIEMENS

Results – Scalability

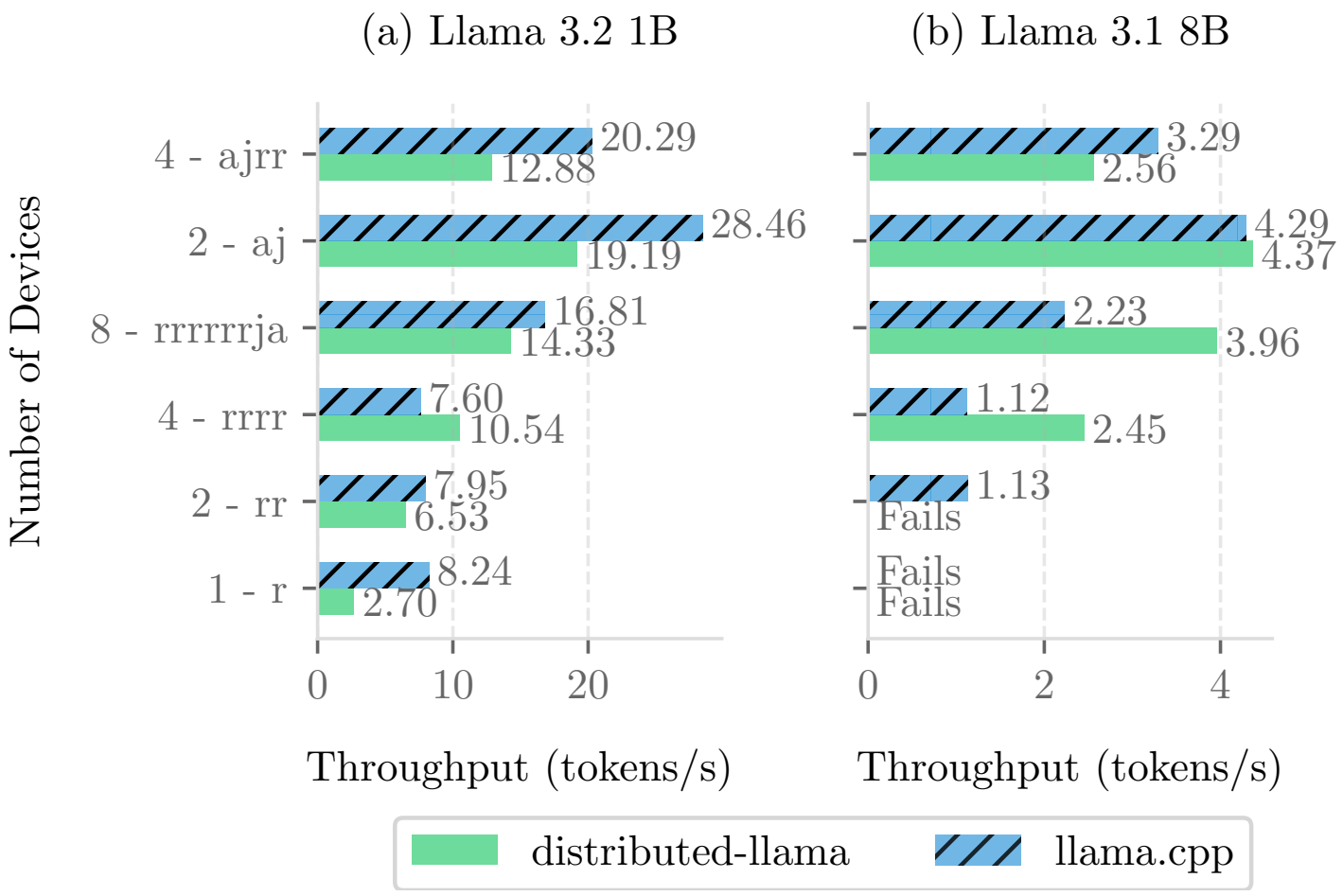


Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
  - More efficient RAM usage
  - Lower throughput for larger models
- `distributed-llama`:
  - More devices → Higher throughput
  - Higher throughput for larger models
- Both:
  - Better devices → Higher throughput
  - Less throughput including worse devices

SIEMENS

Results – Scalability

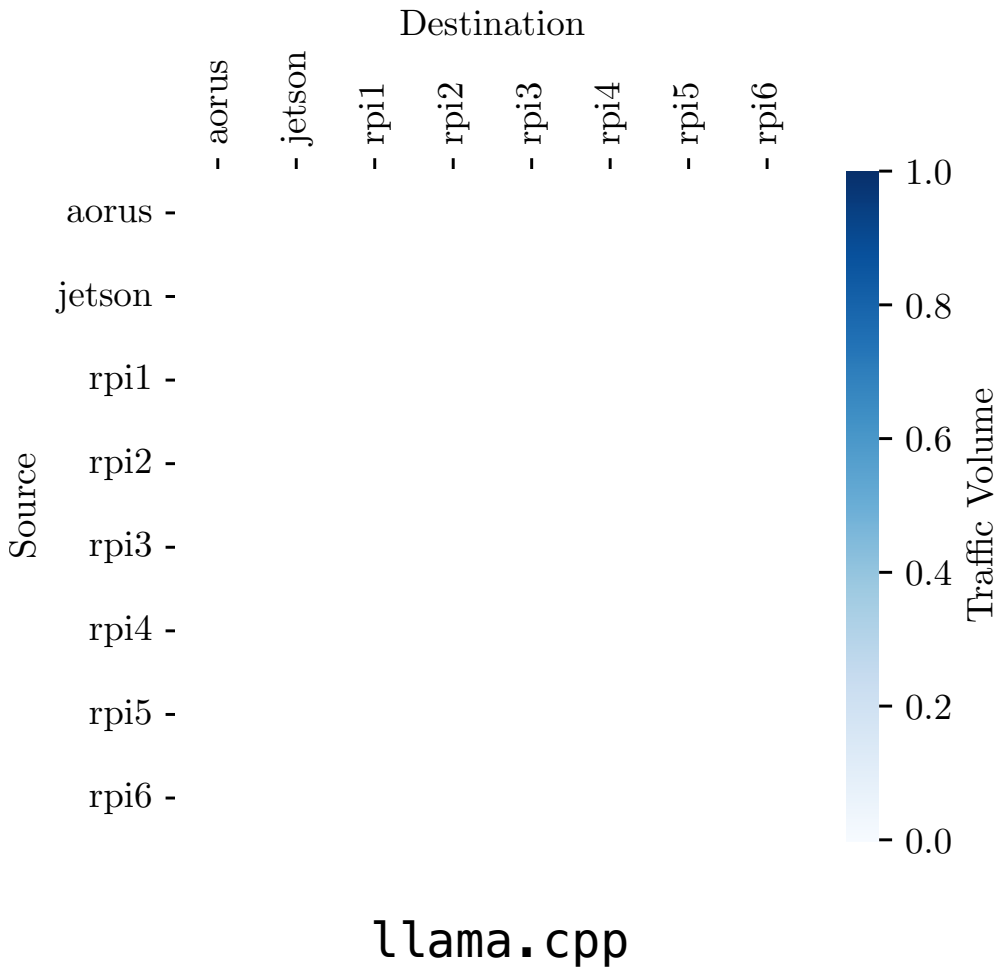


Results:

- `llama.cpp`:
  - More devices → Same throughput
  - Better devices → Higher throughput
  - More efficient RAM usage
  - Lower throughput for larger models
- `distributed-llama`:
  - More devices → Higher throughput
  - Higher throughput for larger models
- Both:
  - Better devices → Higher throughput
  - Less throughput including worse devices

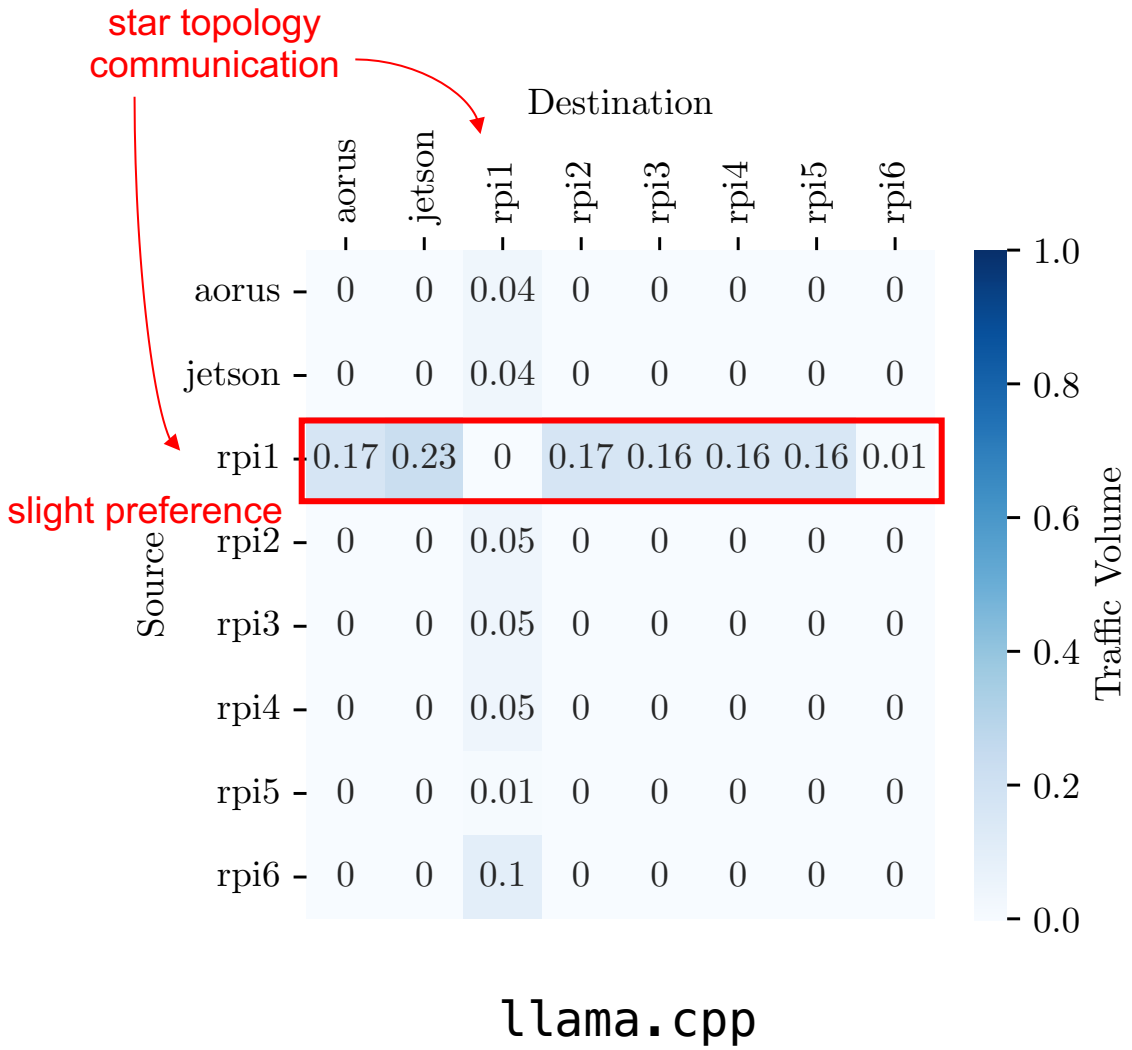
SIEMENS

Results – Traffic Volume



SIEMENS

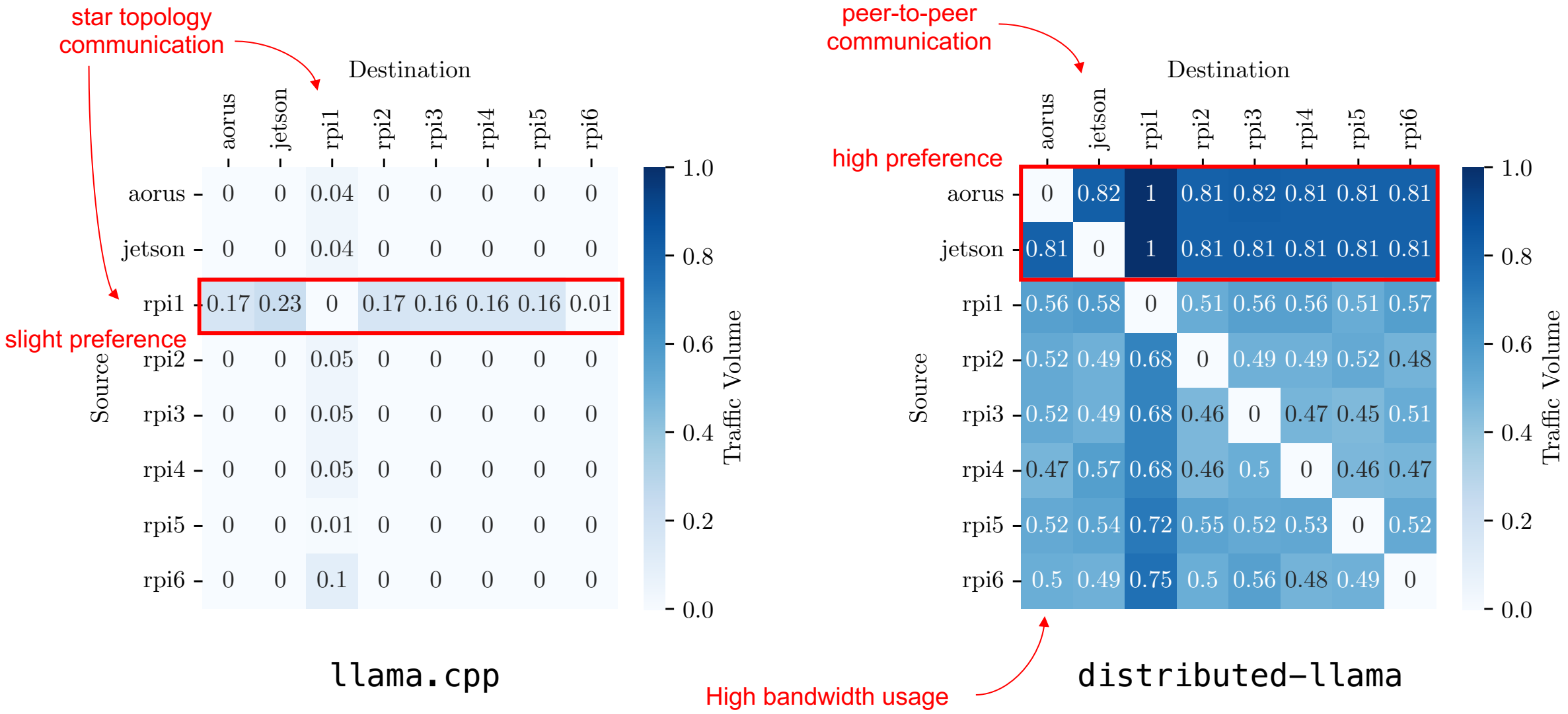
Results – Traffic Volume



`llama.cpp`

SIEMENS

Results – Traffic Volume



SIEMENS

Takeaways – It depends…

Choose your distributed LLM inference framework:

1. Heterogeneous or homogeneous cluster
2. High or low number of devices
3. Available Bandwidth
4. Inference throughput

- **`llama.cpp`:**
  - Few powerful, homogeneous devices
  - Lower bandwidth usage

- **`distributed-llama`:**
  - Maximum horizontal scalability
  - Many devices (even heterogeneous)
  - Higher throughput for large models

Machine operator is happy

John is happy

Contact

Philippe Buschmann
PhD Candidate

Siemens AG
Foundational Technologies
Research & Predevelopment
Industrial Networks & Wireless Germany
FT RPD CED INW-DE
Friedrich-Ludwig-Bauer-Str. 3
85748 Garching, Germany

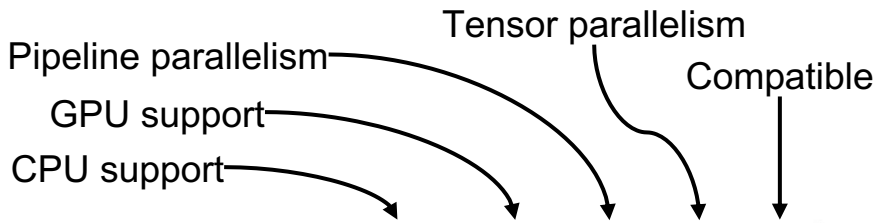
E-mail: philippe.buschmann@siemens.com

SIEMENS

Appendix

SIEMENS

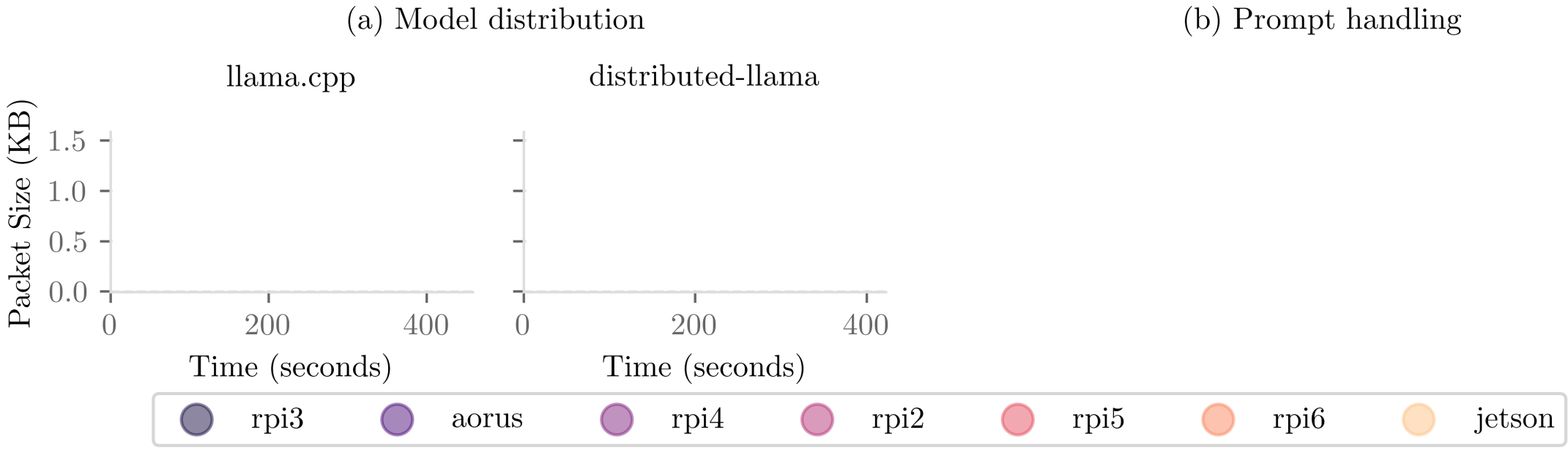
Potential and Selected Frameworks



Framework	CPU	GPU	PP	TP	C
Petals [16]	✗	✓	✓	✗	✗
LLM-PQ [35]	✗	✓	✓	✗	✗
HexGen [21]	✗	✓	✓	✓	✗
vLLM [13]	✓[1]	✓	✓	✓	✗
exo [3]	✓	✓	✓	✗	✗
prima.cpp [23]	✓	✓	✓	✗	✓
llama.cpp [5]	✓	✓	?	?	✓
distributed-llama [31]	✓	✓[2]	✗	✓	✓

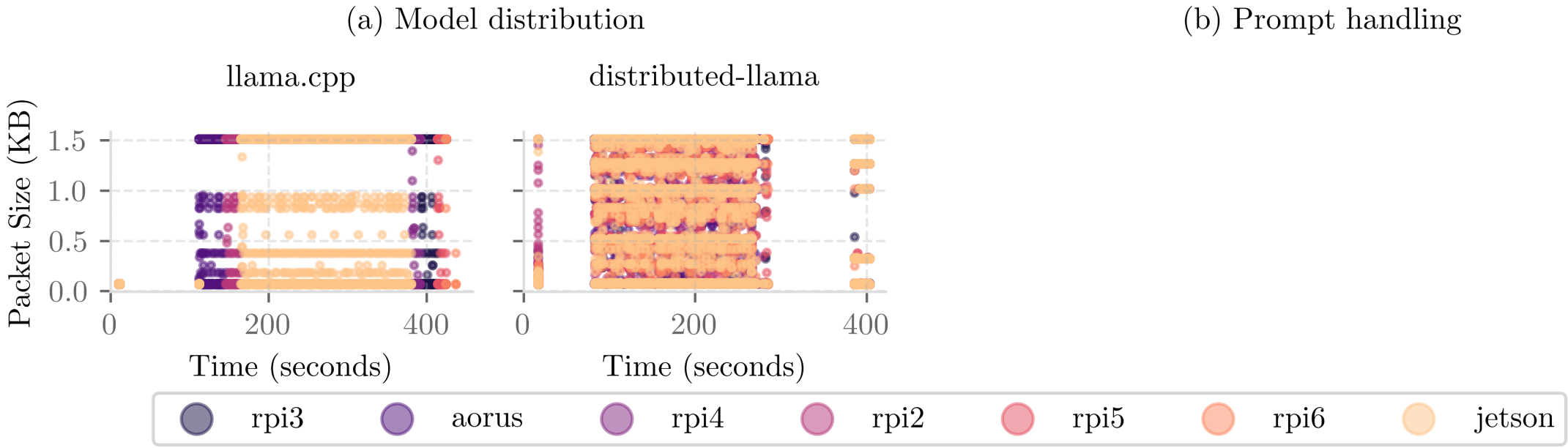
[1] Not available as pre-built
[2] Vulkan (experimental)

Results – Traffic Patterns



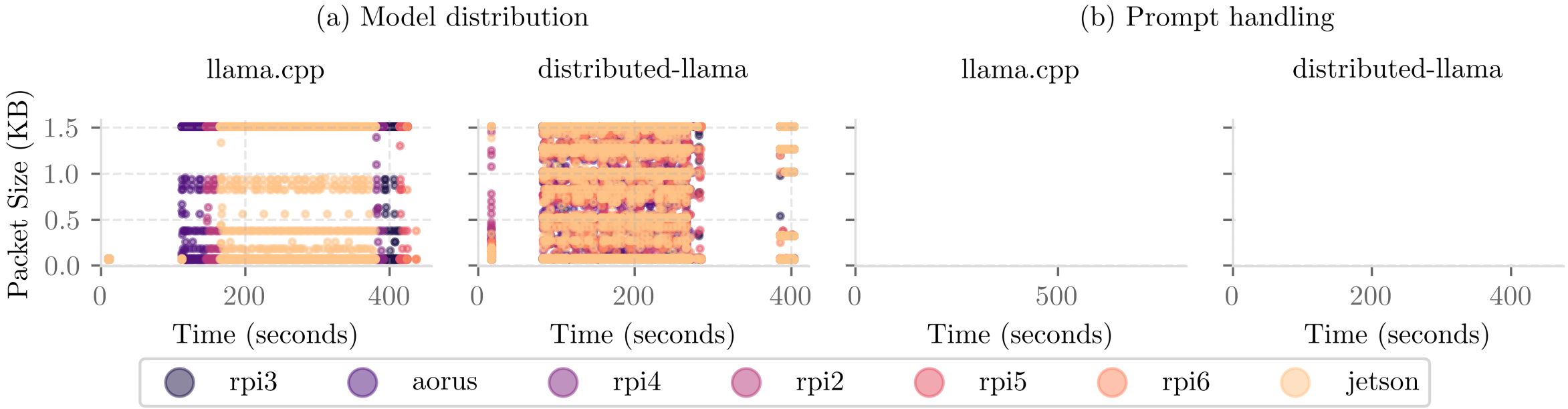
SIEMENS

Results – Traffic Patterns



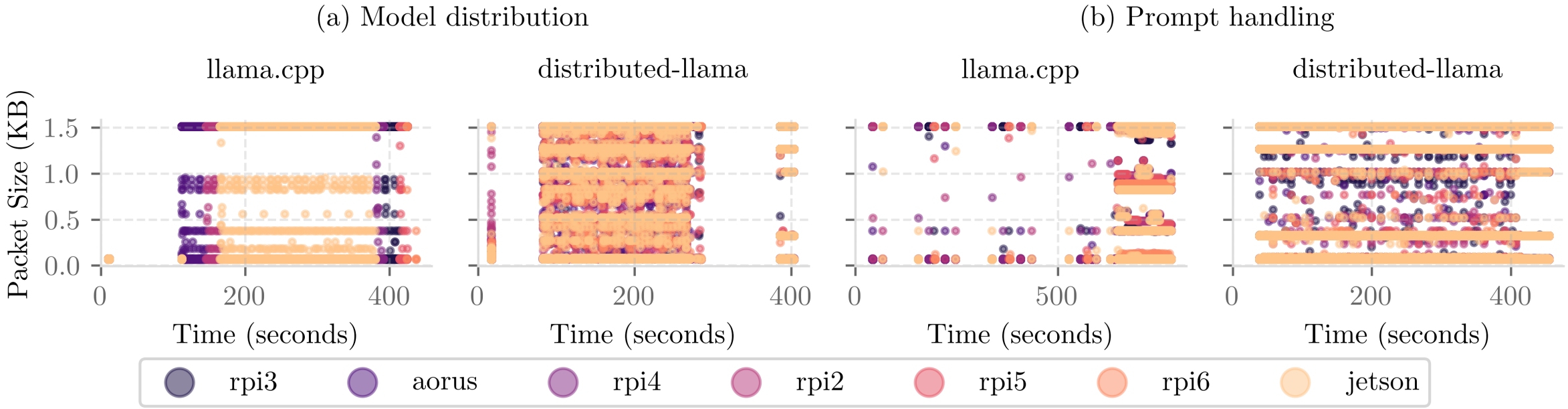
SIEMENS

Results – Traffic Patterns



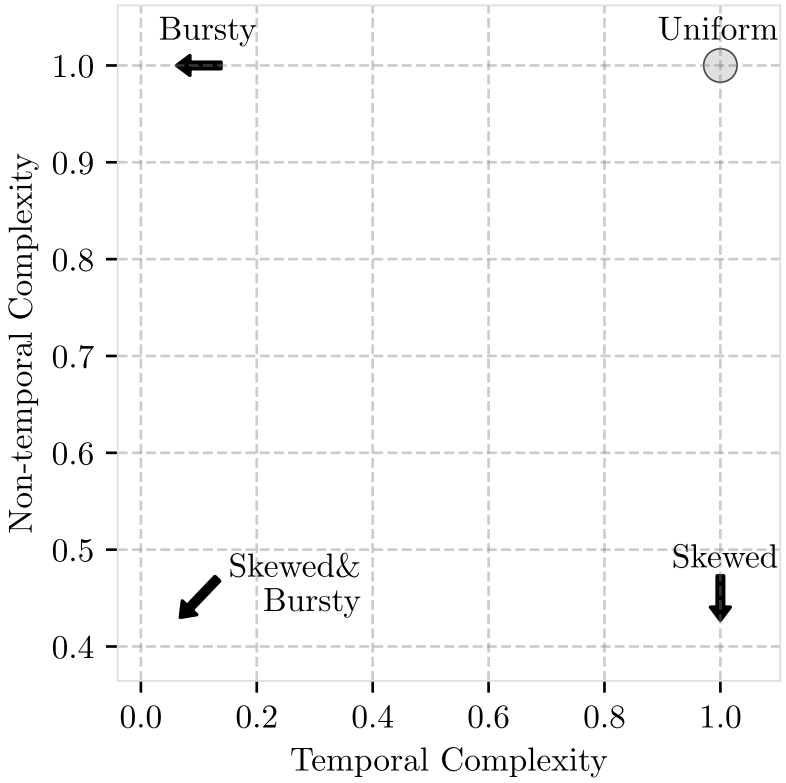
SIEMENS

Results – Traffic Patterns

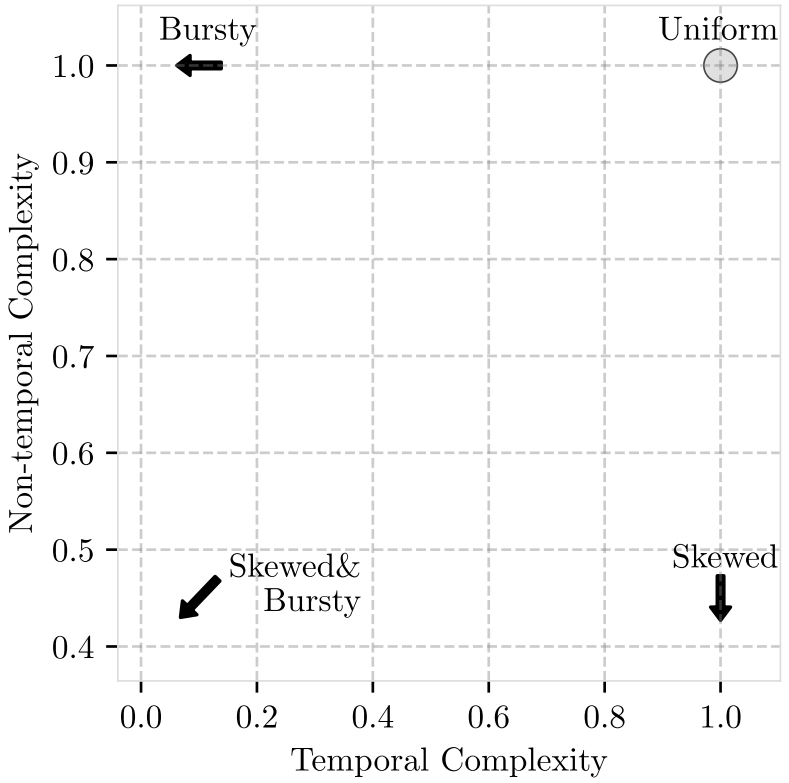


SIEMENS

Results – Traffic Structure



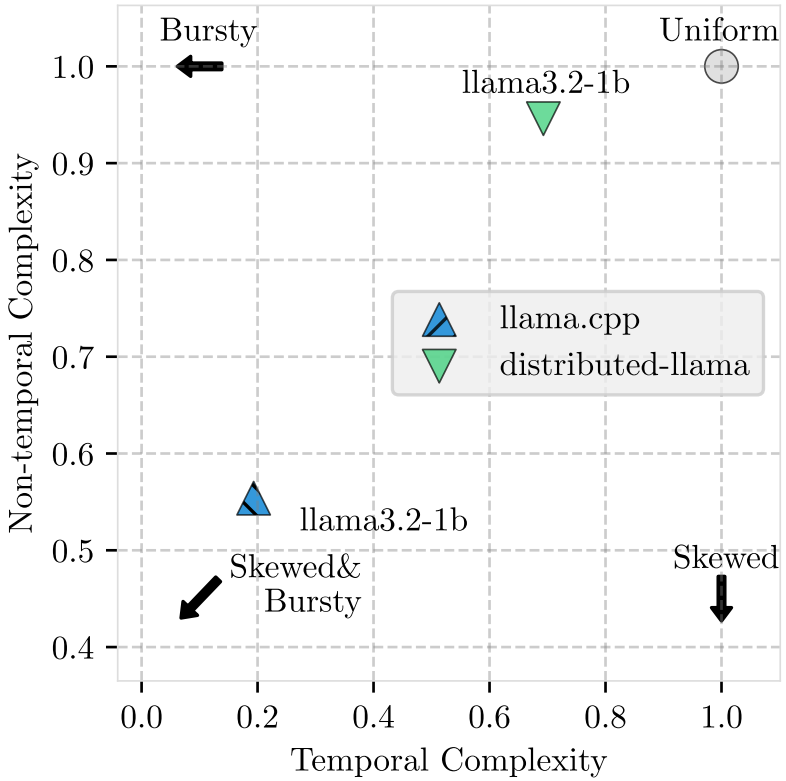
Model Distribution



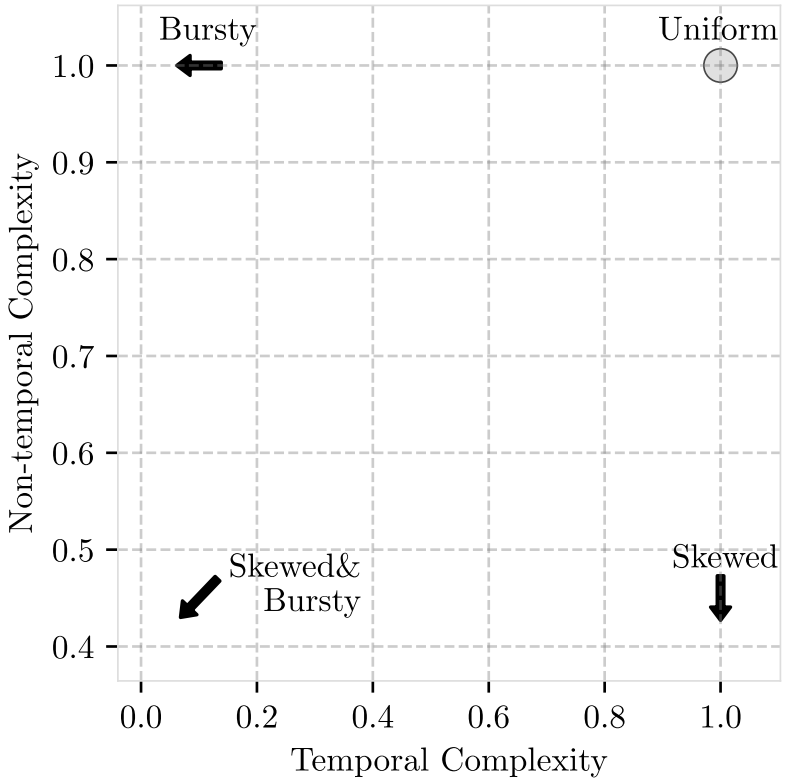
Prompt Handling

SIEMENS

Results – Traffic Structure



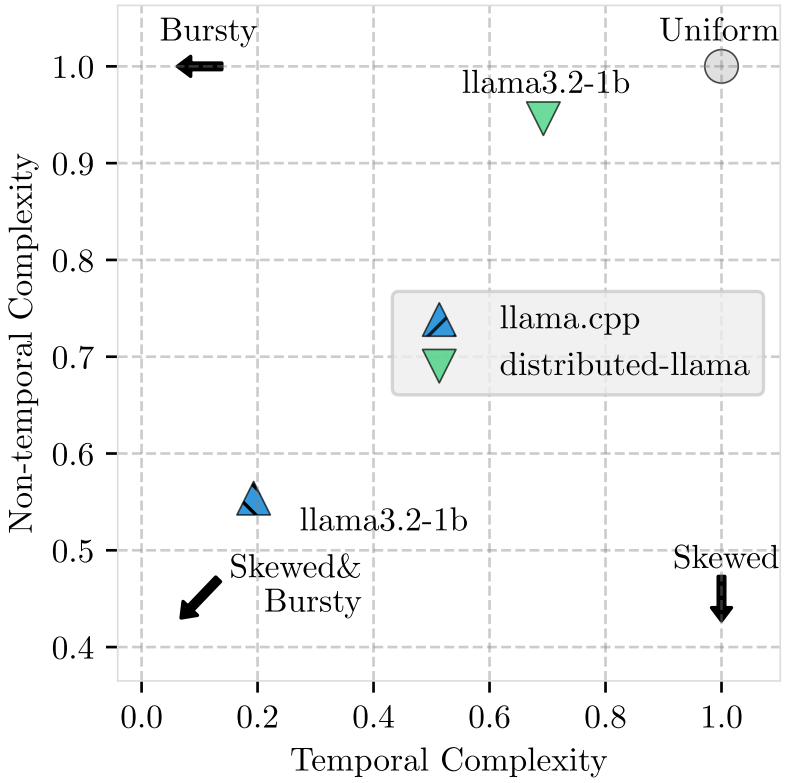
Model Distribution



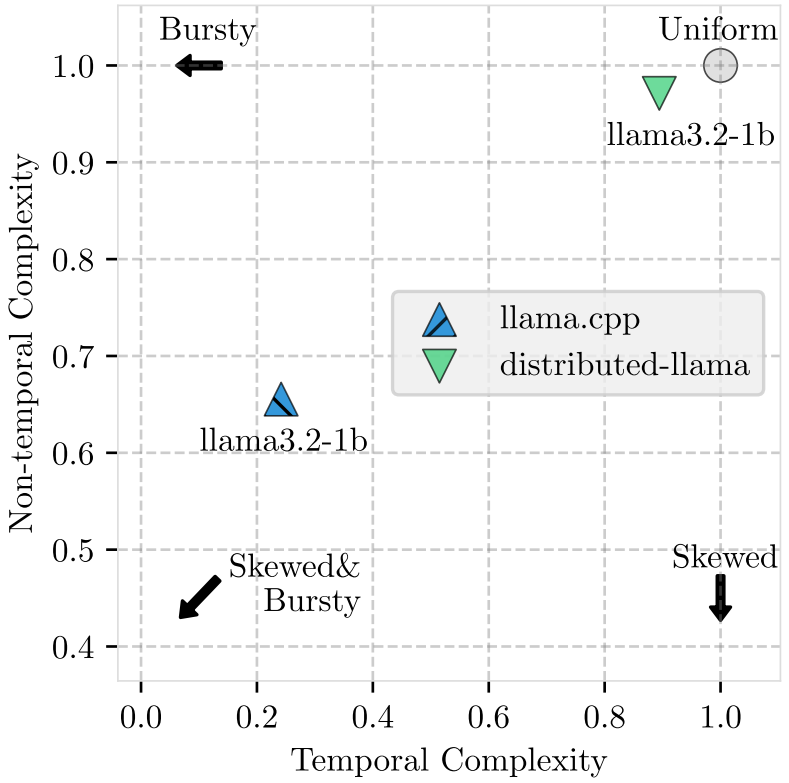
Prompt Handling

SIEMENS

Results – Traffic Structure



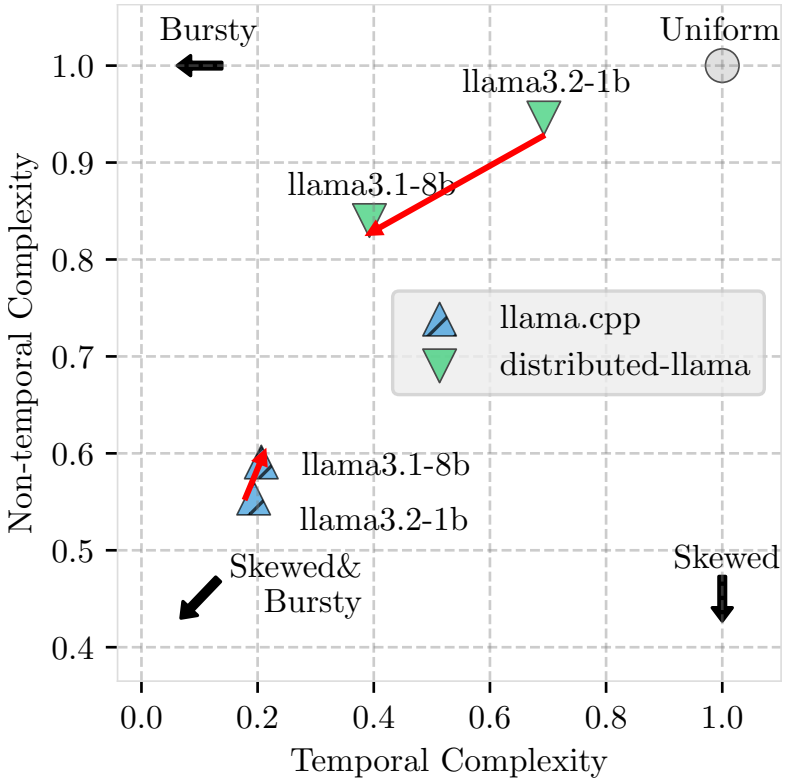
Model Distribution



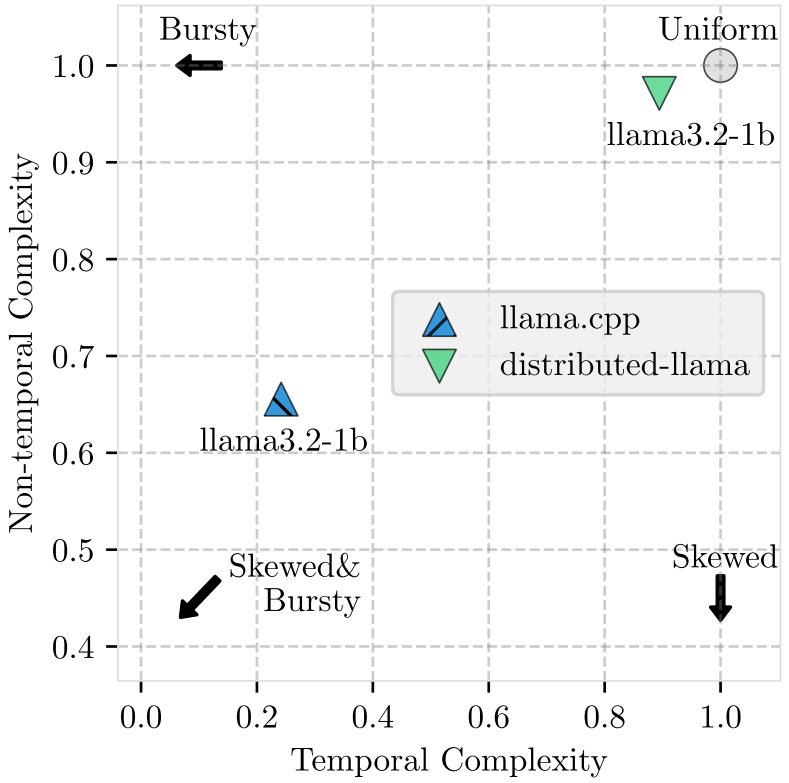
Prompt Handling

SIEMENS

Results – Traffic Structure



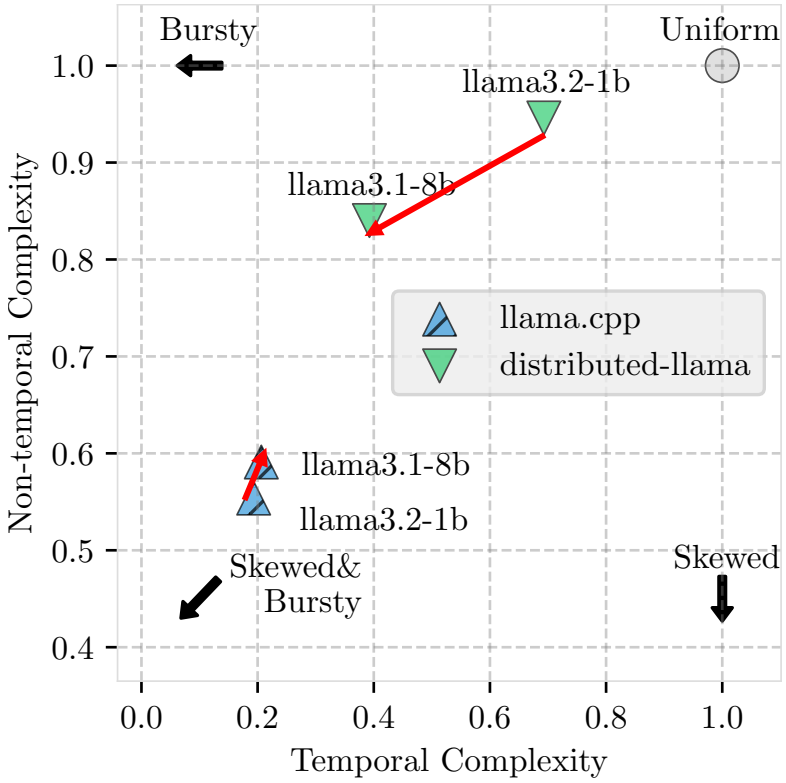
Model Distribution



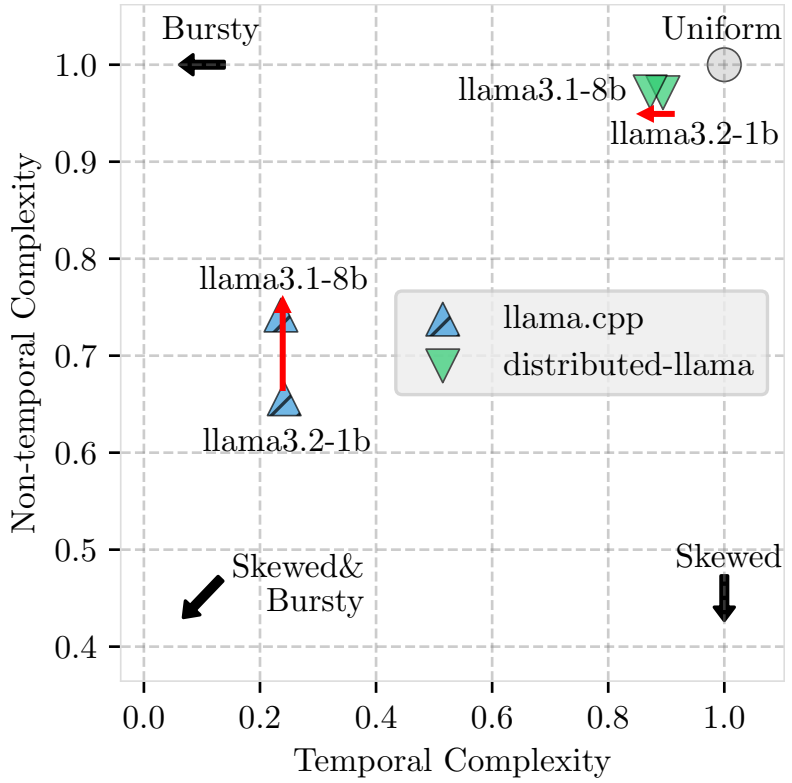
Prompt Handling

SIEMENS

Results – Traffic Structure



Model Distribution



Prompt Handling

SIEMENS