SIMULATING LLM ECOSYSTEMS

FROM **REFERENCE ARCHITECTUREs** OF LLM ECOSYSTEMS TO **SIMULATING** PERFORMANCE, SUSTAINABILITY, EFFICIENCY

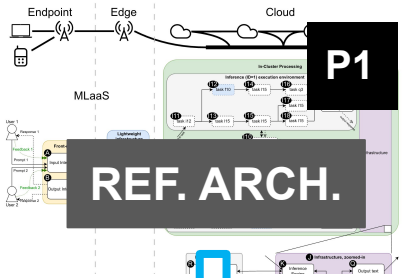
@Large Research
Massivizing Computer Systems

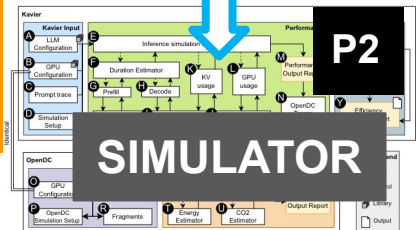
These slides:
bit.ly/radu-bsc

Radu Nicolae
@VU Amsterdam
@Large Research

@rnicolae,
mail@radu-nicolae.com

Towards digital twinning LLM Ecosystems
= **ref. arch.** + **simulator** + monitoring, datagen + ODA +goal-oriented infr. steering





LLM(s)
(ecosystems under inference)

LLMs are widely adopted everywhere

LLMs are being widely adopted, worldwide, amongst academia, government, and industry.

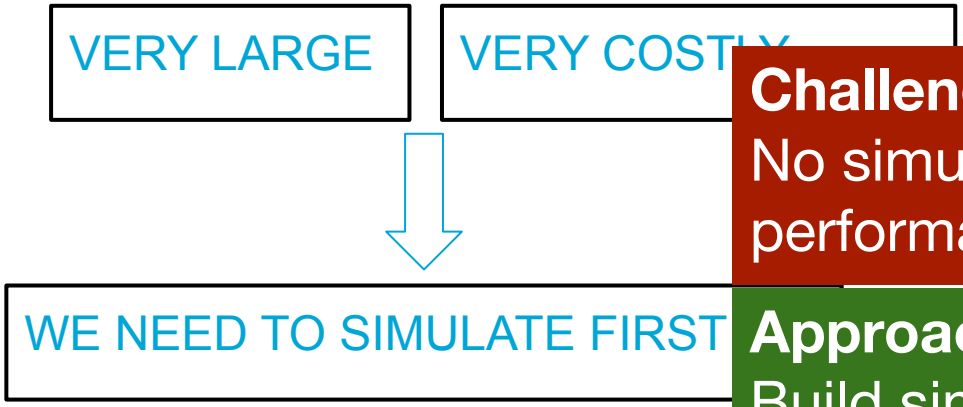
(Microsoft)scaling at all costs

Source: The guardian (online)

SUSTAINABILITY	FINANCIAL	PERFORMANCE
Inference GPT: 1 GWh daily Equivalent to ~40,000 dutch	Inference GPT: $700k daily (operational costs)	Modern day moore's law

Deploying, monitoring, and scaling LLMs

LLMs



Simulators enable performing large-scale experiments in a time and cost efficient way.

Challenge:
No simulator for predicting LLM performance & sustainability

P2

Approach:
Build sim. adhering to ref. arch.

Challenge:
No (good) reference architecture of LLM ecosys under inference

P1

Tackling:
Design+validate ref arch.

P1
02

A REFERENCE ARCHITECTURE FOR LLM ECOSYSTEMS UNDER INFERENCE

DESIGN [1/4]

C1

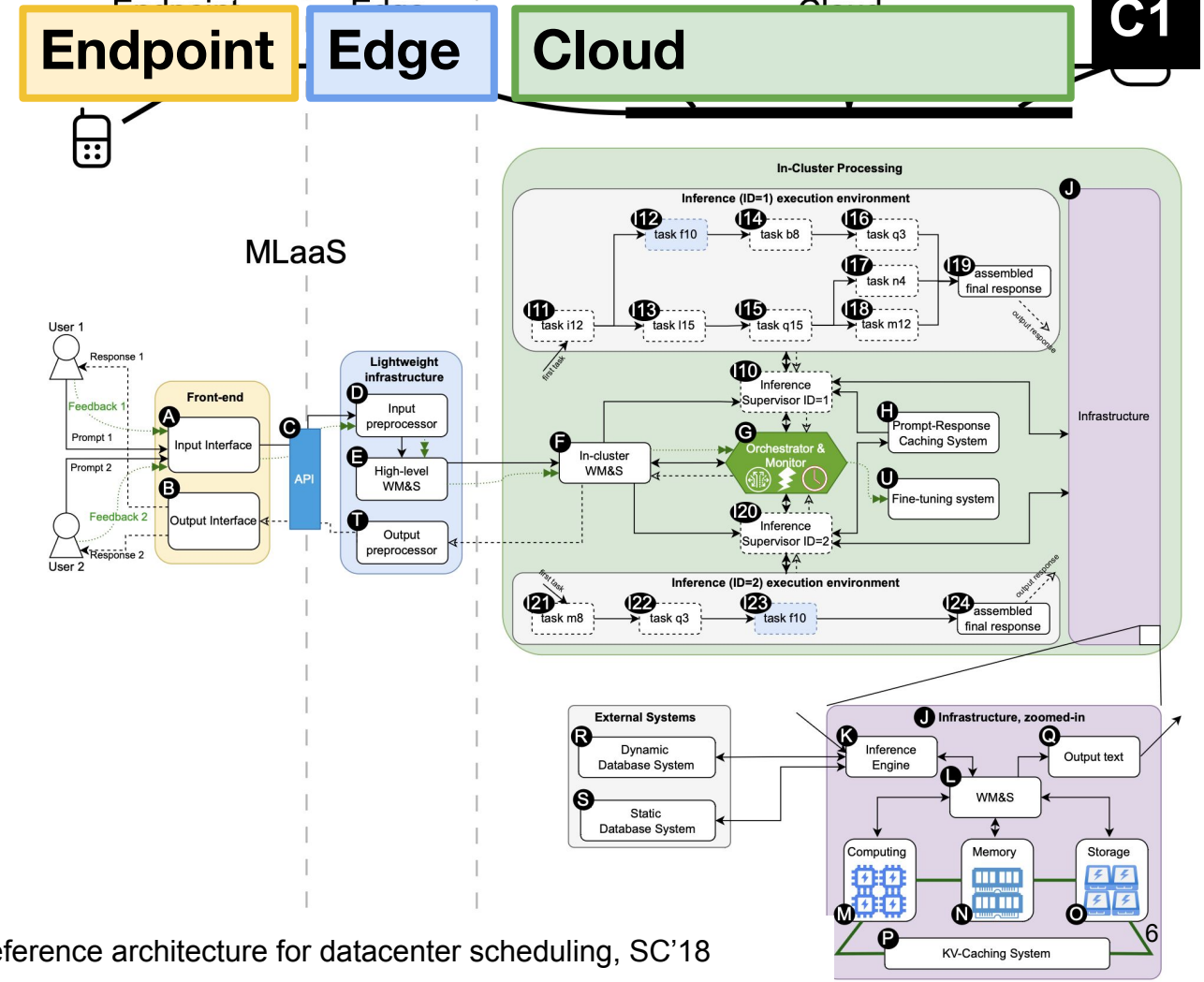
Design requirements[1]

- Validity
- Usefulness

Design principles

- **DP1** Heterogeneity
- **DP2** End-to-end workflow
- **DP3** Branching prompts
- **DP4** Multi-user ecosys
- **DP5** Inference stages

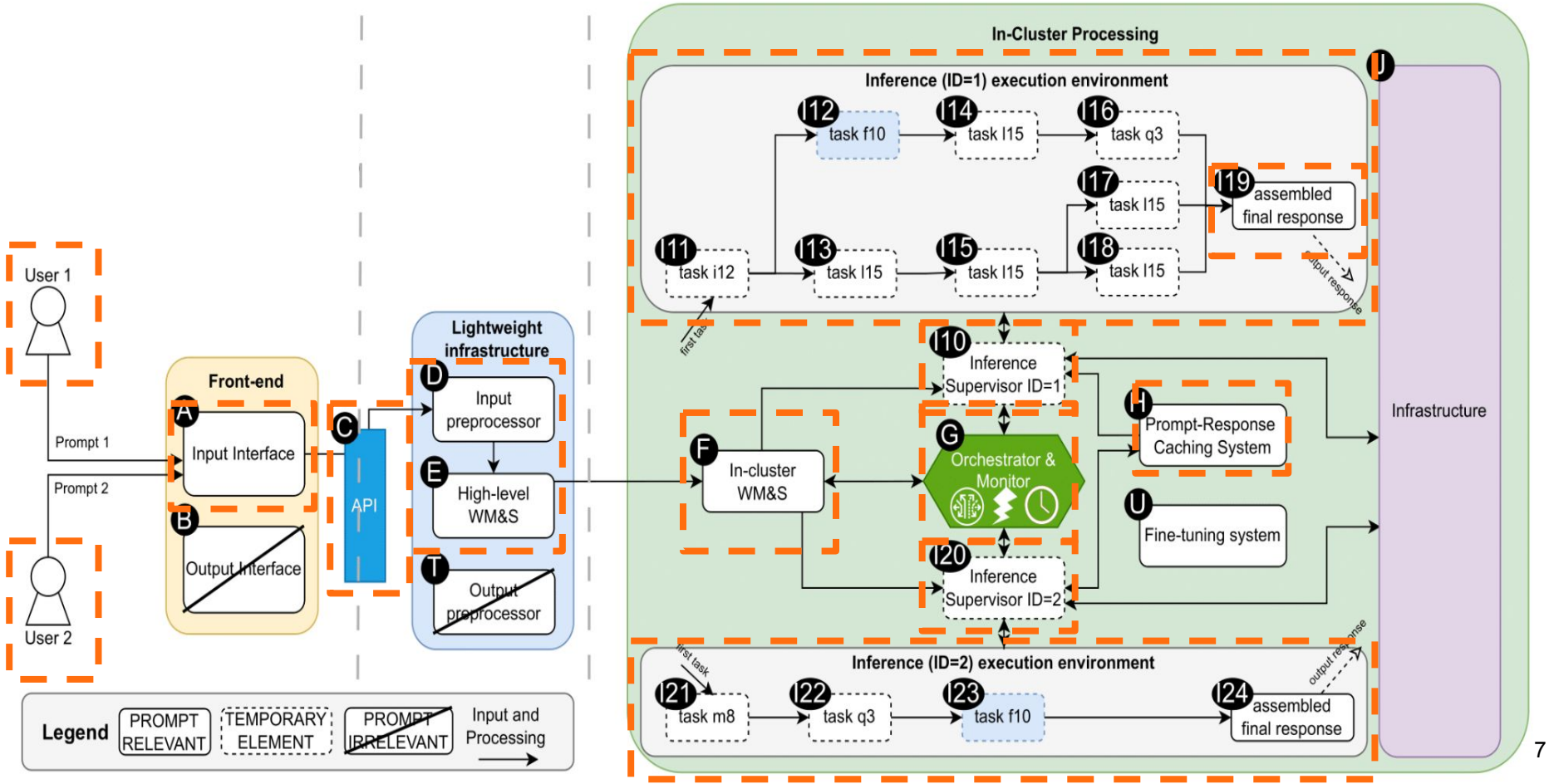
& more



[1] Andreadis et al., A reference architecture for datacenter scheduling, SC'18

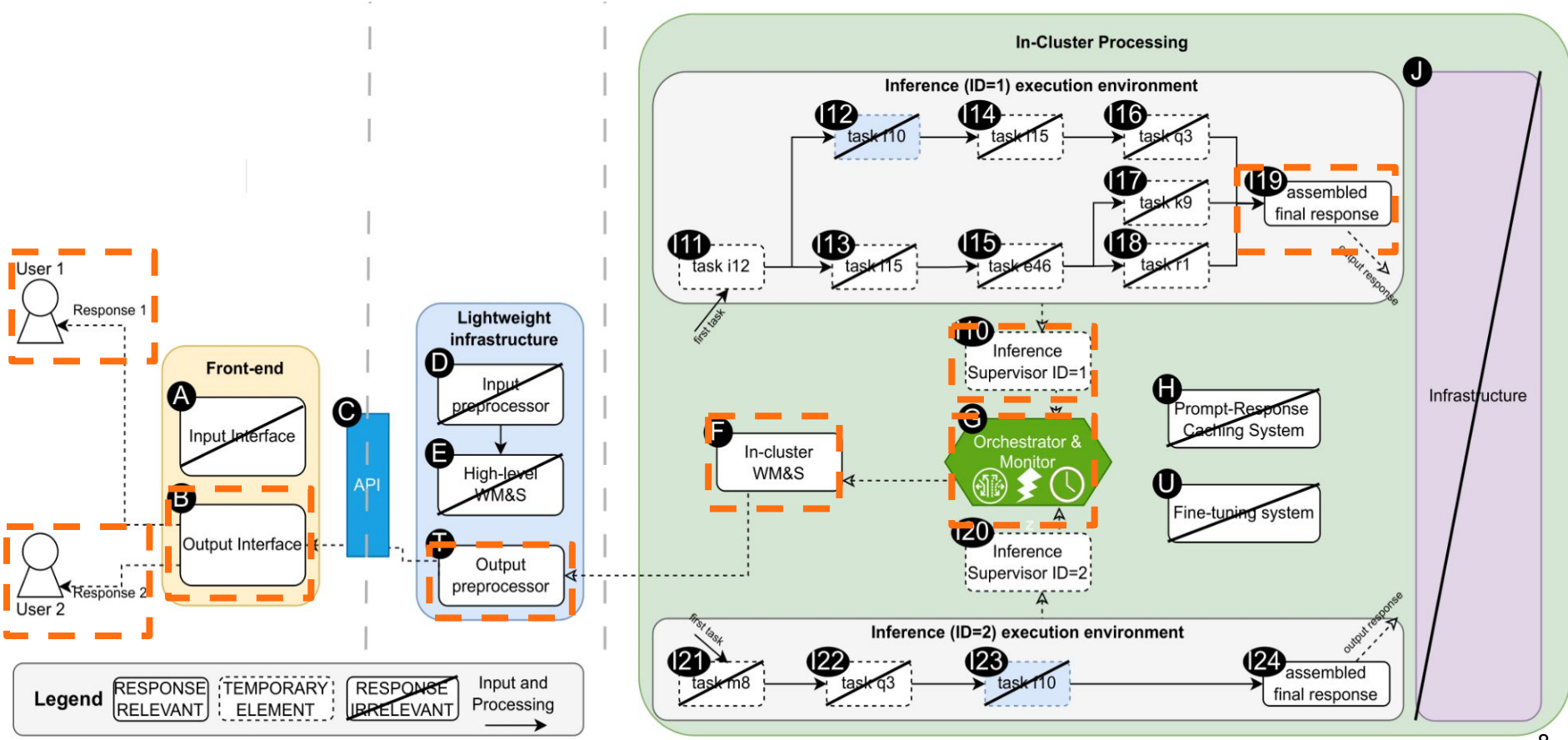
DESIGN - PROMPT WORKFLOW [2/4]

C1



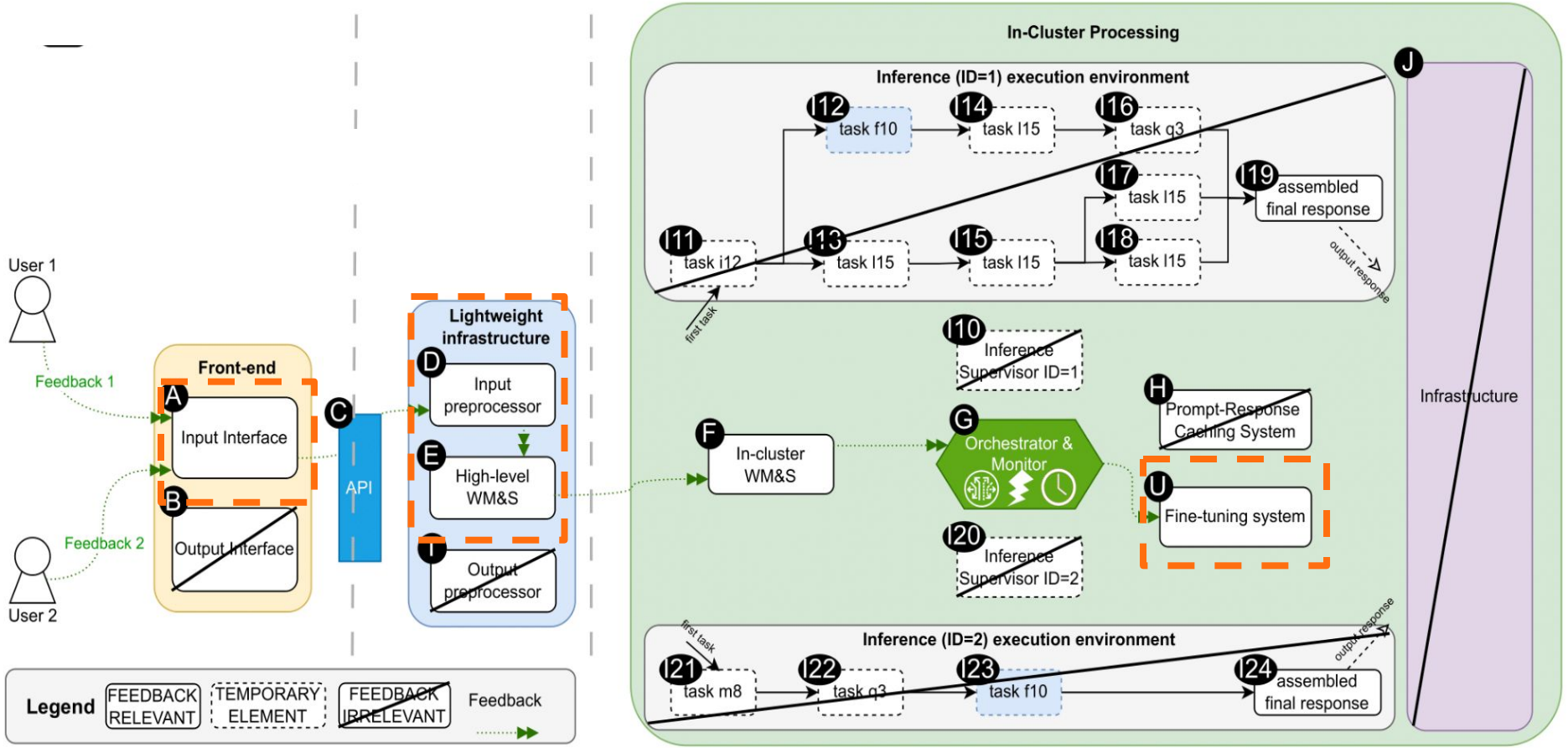
DESIGN - RESPONSE WORKFLOW [3/4]

C1



DESIGN - FEEDBACK WORKFLOW [4/4]

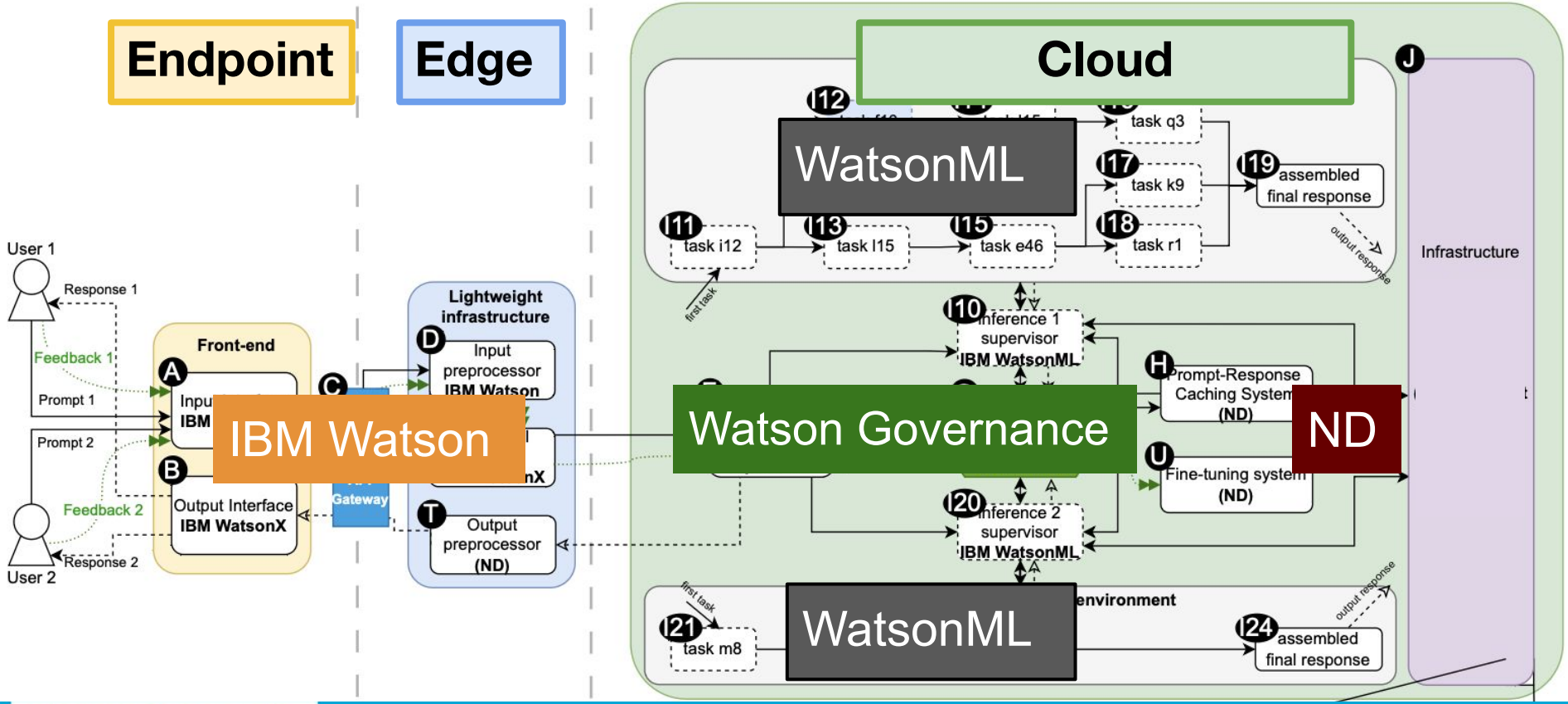
C1



VALIDATION

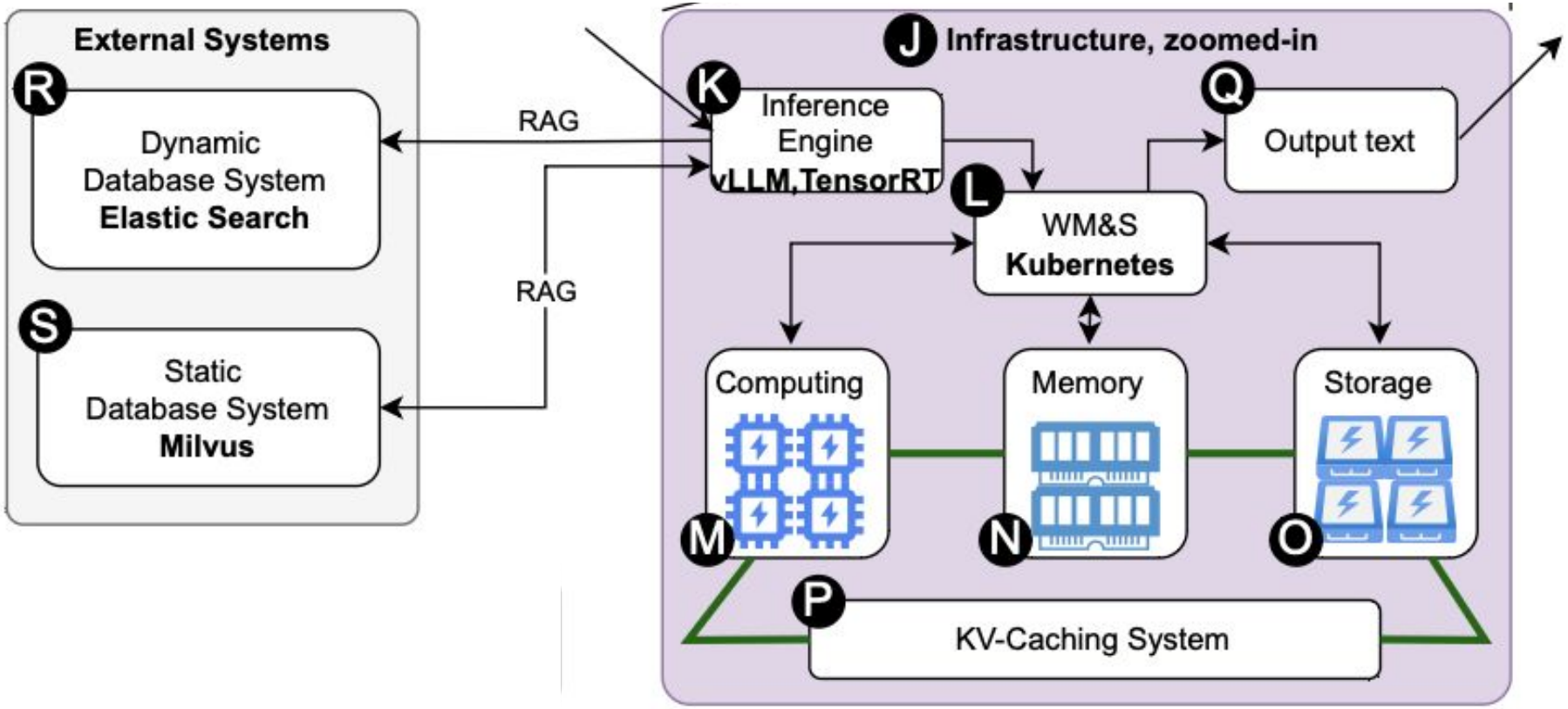
VALIDATION - IBM [1/2]

C2



VALIDATION - IBM [2/2]

C2



TAKE-HOME MESSAGES (REF. ARCHI.)

SIMULATING LLM ECOSYSTEMS

FROM **REFERENCE ARCHITECTUREs** OF LLM ECOSYSTEMS TO
SIMULATING PERFORMANCE, SUSTAINABILITY, EFFICIENCY

@Large Research
Massivizing Computer Systems

Radu Nicolae
@VU Amsterdam
@Large Research

@rnicolae,
mail@radu-nicolae.co

P1

C1: Propose reference architecture

C2: Validate ref. arch. (x3) (-> x4?)

P2

C3: Design LLM inference simulator

C4: Impl. sim, integrate with OpenDC

C5: Trace vLLM deployments

C6: Validate simulator *(wip)*

P2

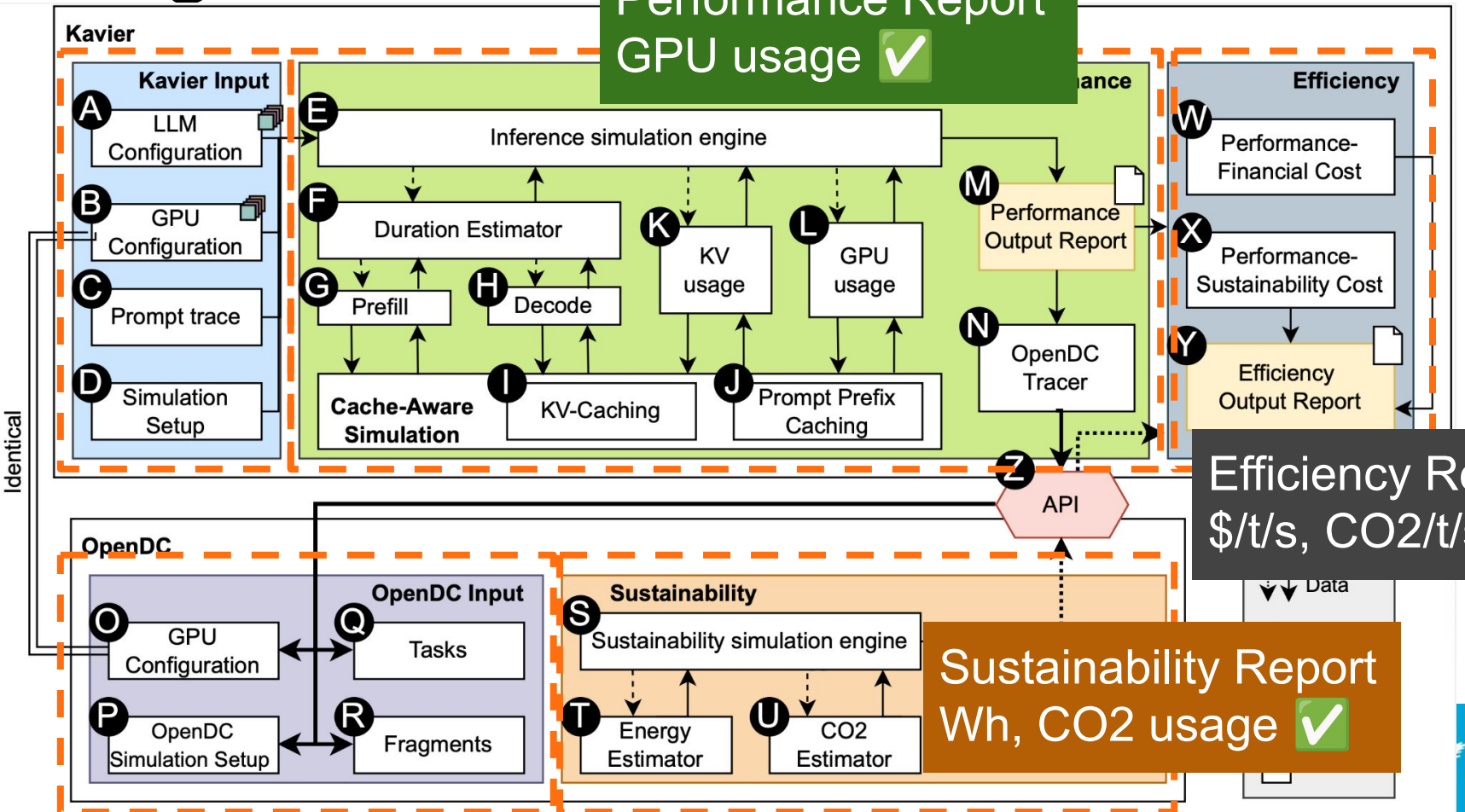
01

KAVIER:

A SIMULATOR FOR LLM ECOSYSTEMS UNDER INFERENCE (DESIGN, IMPLEMENTATION, INTEGRATION)

Design of Kavier

Performance Report
GPU usage ✅



Efficiency Report
$/t/s, CO2/t/s ✅

Sustainability Report
Wh, CO2 usage ✅

C4

P2

01

KAVIER:

IMPLEMENTED A PROTOTYPE, INTEGRATED WITH OPENDC, OPEN-SOURCE & OPEN-SCIENCE

TRACER

LLM Ecosystem Tracer

C5

Challenge: no traces decode/prefill time per decode/prefill length

Approach: accessed clusters, deployed vLLM, measured performance

NETWORK INSTITUTE

SURF

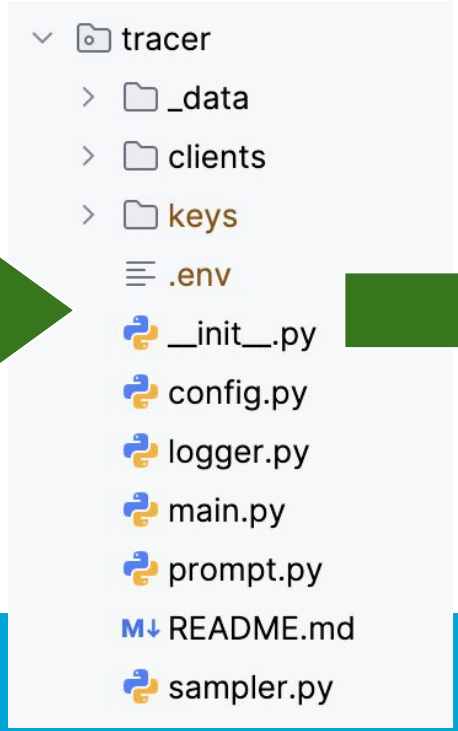
vLLM

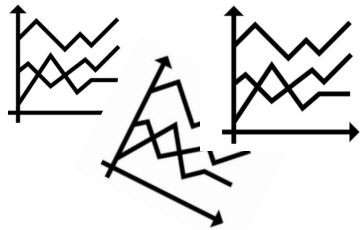
A10

DAS 6

vLLM

A4000, A6000, A100 ++





Traces released open-source & open-science

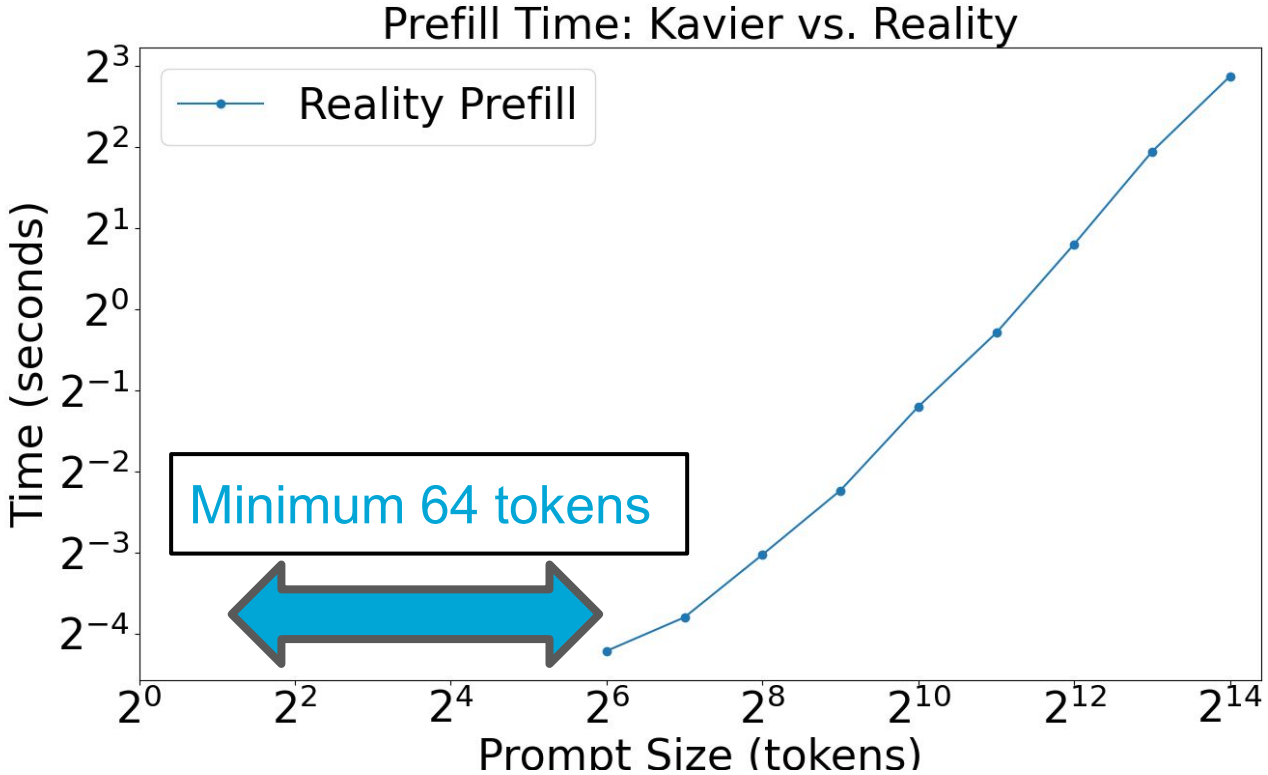
@Large Research
Massivizing Computer Systems

VALIDATION

Validation - Prefill [1/4]

Experiment setup:

- SURF
- A10
- vLLM

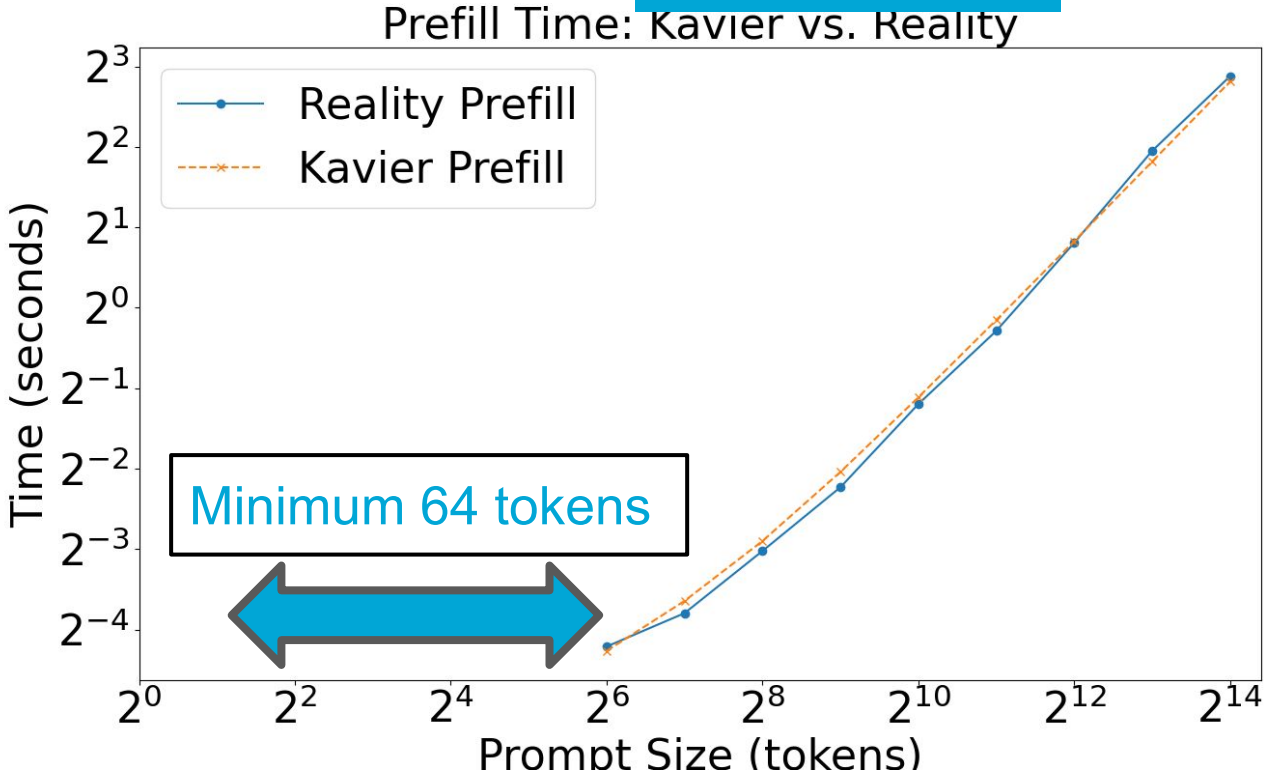


Validation - Prefill [2/4]

MAPE: 7.58%

Experiment setup:

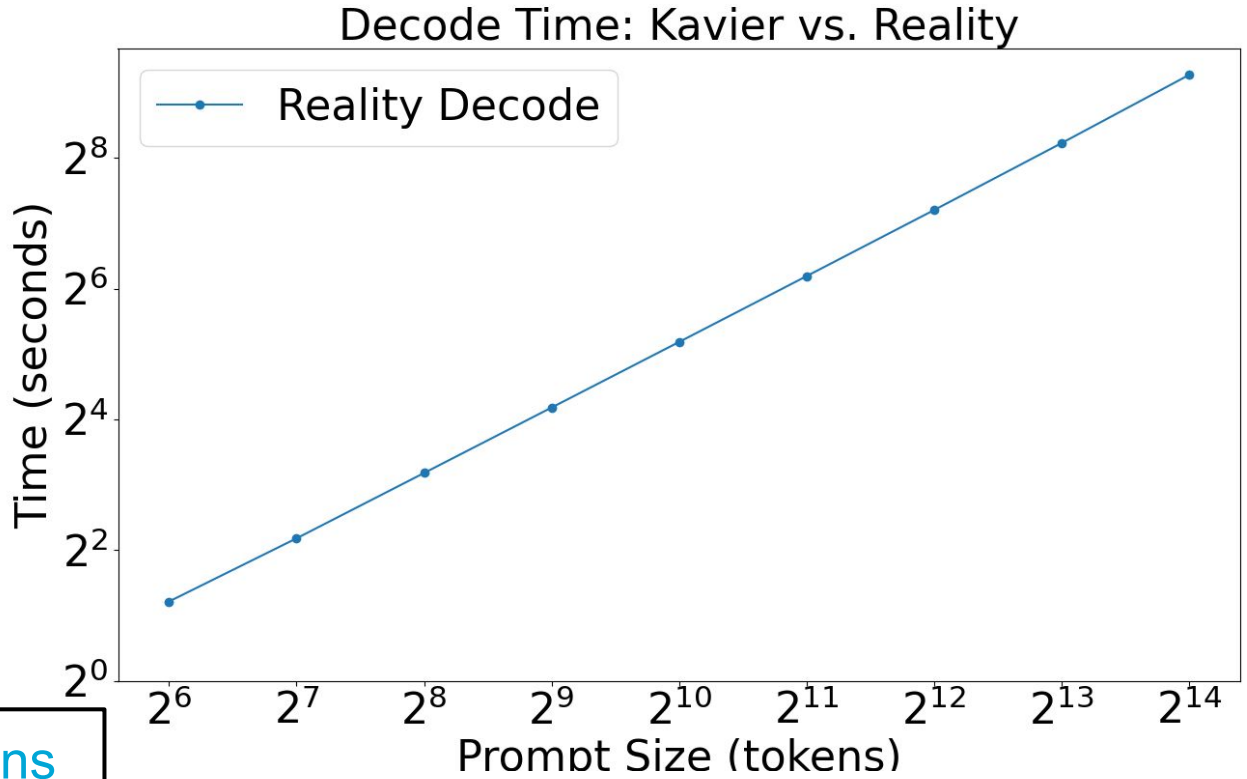
- SURF
- A10
- vLLM



Validation - Decode [3/4]

Experiment setup:

- SURF
- A10
- vLLM



Starting at 64 tokens

Minimum prompt size 64 tokens;
this illustration (horizontal-axis) for better visual purposes.

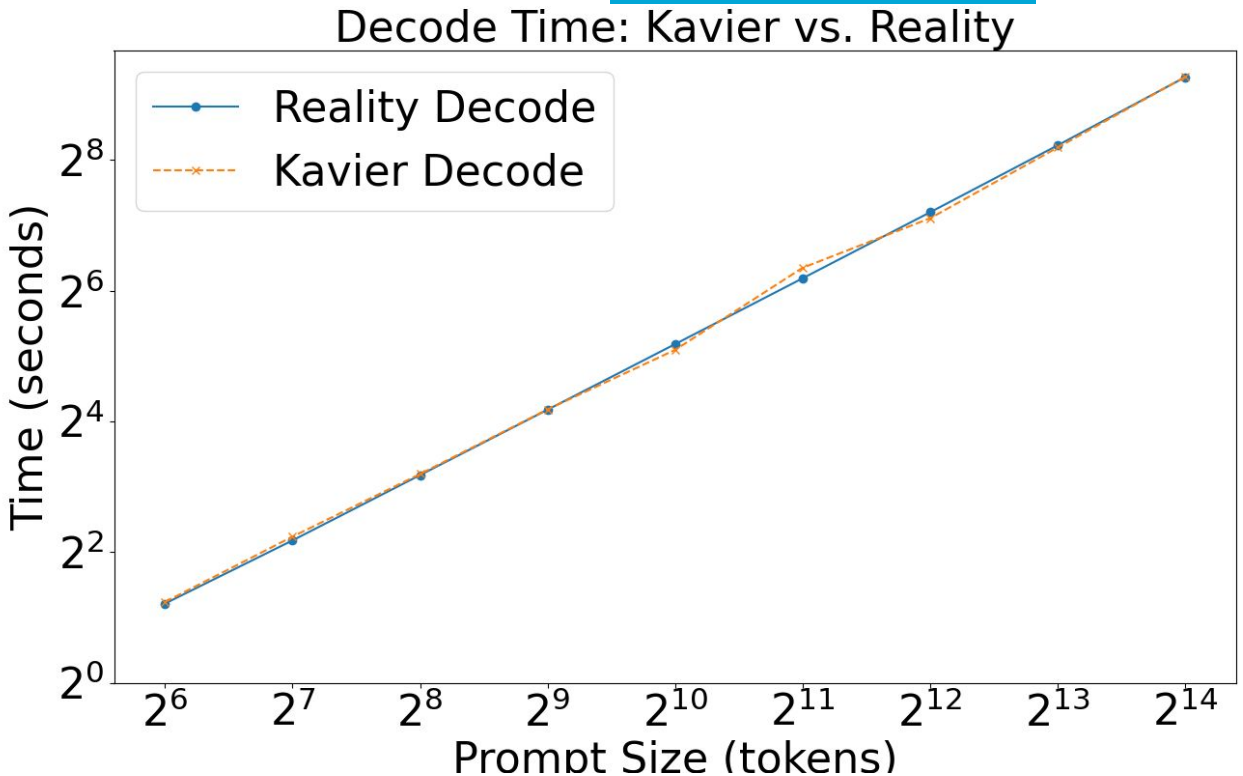
Validation - Decode [4/4]

C6

MAPE: 3.75%

Experiment setup:

- SURF
- A10
- vLLM



SIMULATING LLM ECOSYSTEMS

FROM **REFERENCE ARCHITECTUREs** OF LLM ECOSYSTEMS TO
SIMULATING PERFORMANCE, SUSTAINABILITY, EFFICIENCY

@Large Research
Massivizing Computer Systems

Radu Nicolae
@VU Amsterdam
@Large Research

@rnicolae,
mail@radu-nicolae.co

P1

C1: Propose reference architecture

C2: Validate ref. arch. (x3)

P2

C3: Design LLM inference simulator

C4: Impl. sim, integrate with OpenDC

C5: Trace vLLM deployments

C6: Validate simulator *(wip)*