Scheduler Load tracking updates and improvements

Vincent Guittot
OSPM 3rd April 2017

Agenda

- Status
- Remaining problems
- What's next ?
- Discussion

Status

Load tracking in scheduler

- CFS class
  - PELT geometric serie
- RT class
  - rt_avg
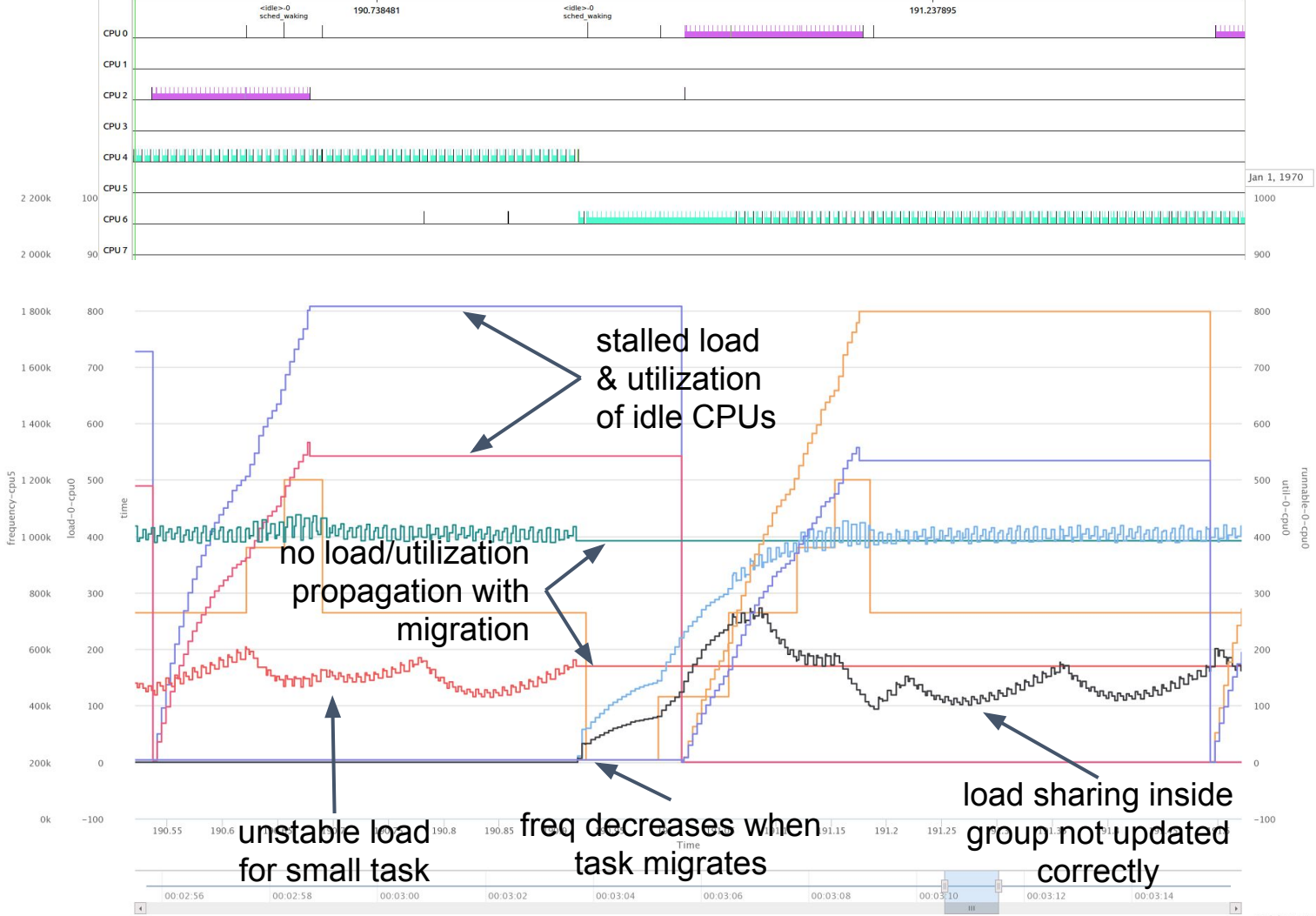- Deadline class
  - Per rq active utilization

Load tracking in scheduler

- Most of the focus has been on CFS and PELT

- RT load tracking
  - rt_avg is not a good metric
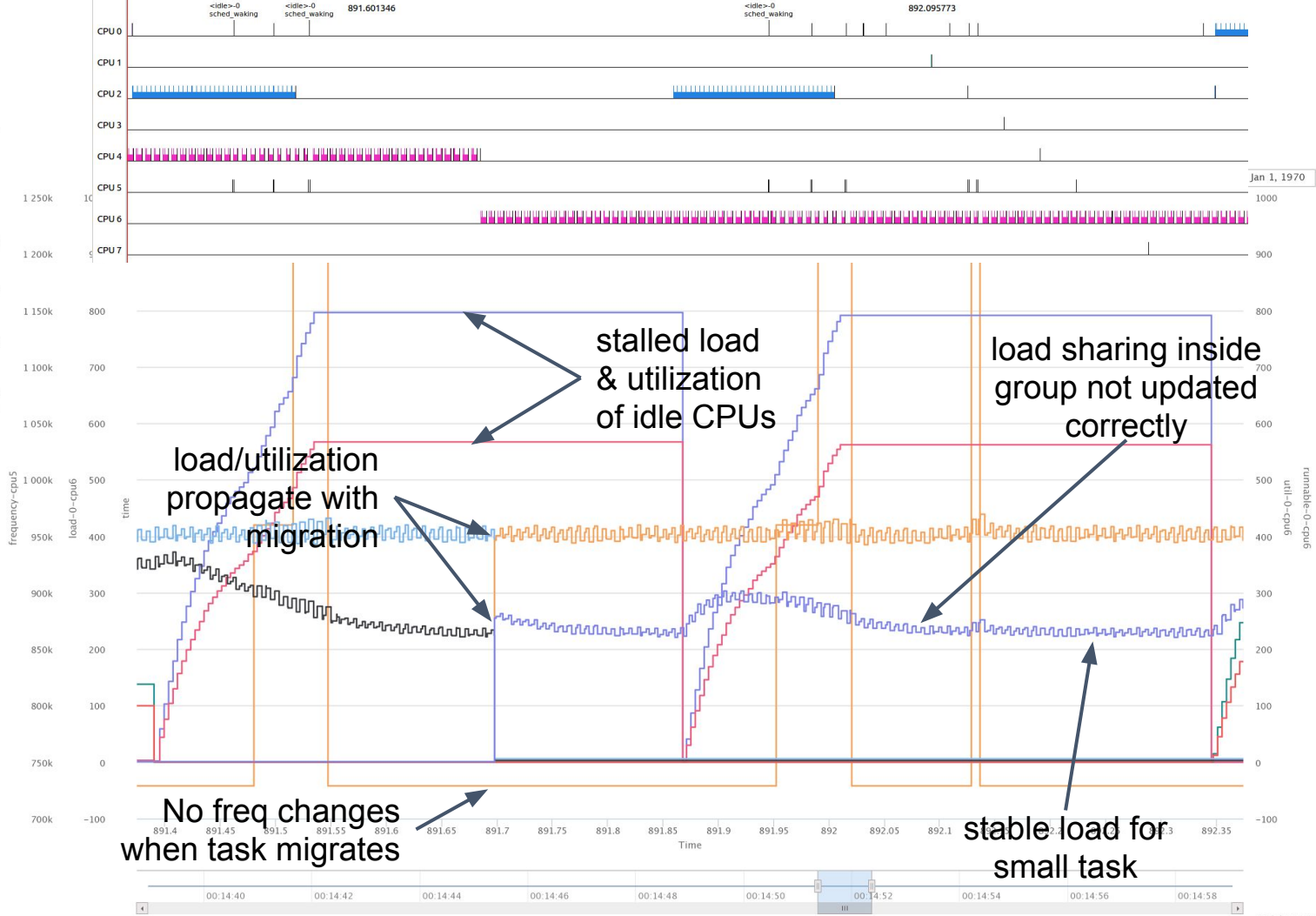
- Deadline utilization
  - See Juri's patchset

v4.9

stalled load & utilization of idle CPUs

no load/utilization propagation with migration

unstable load for small task

freq decreases when task migrates

load sharing inside group not updated correctly

tip

stalled load
& utilization
of idle CPUs

load sharing inside
group not updated
correctly

load/utilization
propagate with
migration

No freq changes
when task migrates

stable load for
small task

Target
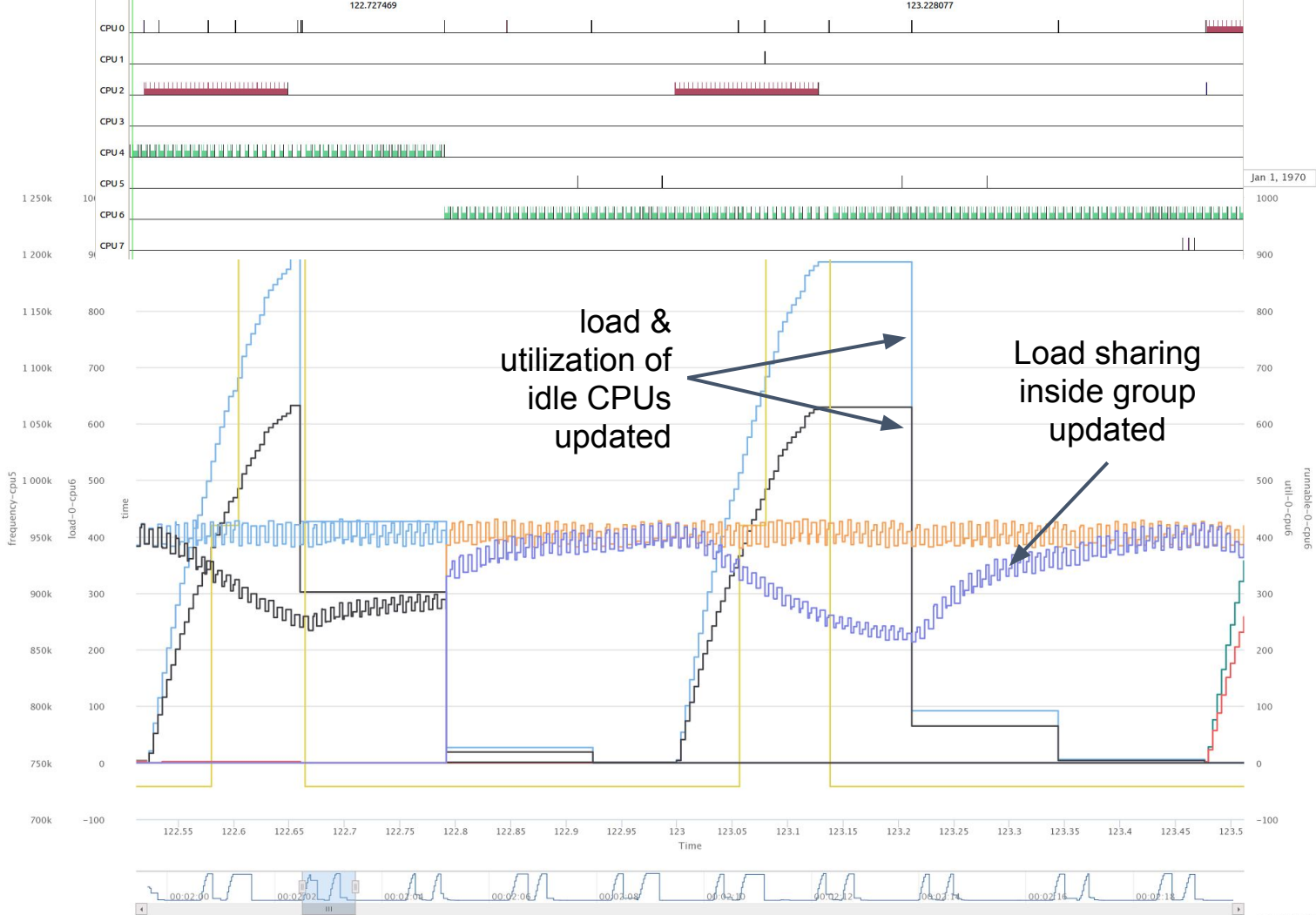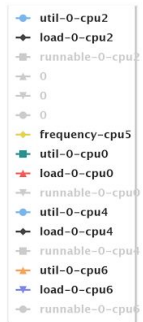
Remaining points

- Frequency invariance
- Update of blocked idle load
- Utilization drop on RT/DL preemption
- Utilization drop on migration

Frequency invariance

Power Consumption or anything else you want to display
Current, Watt or whatever the unit

Utilization and freq scaling w/ new invariance

110ms

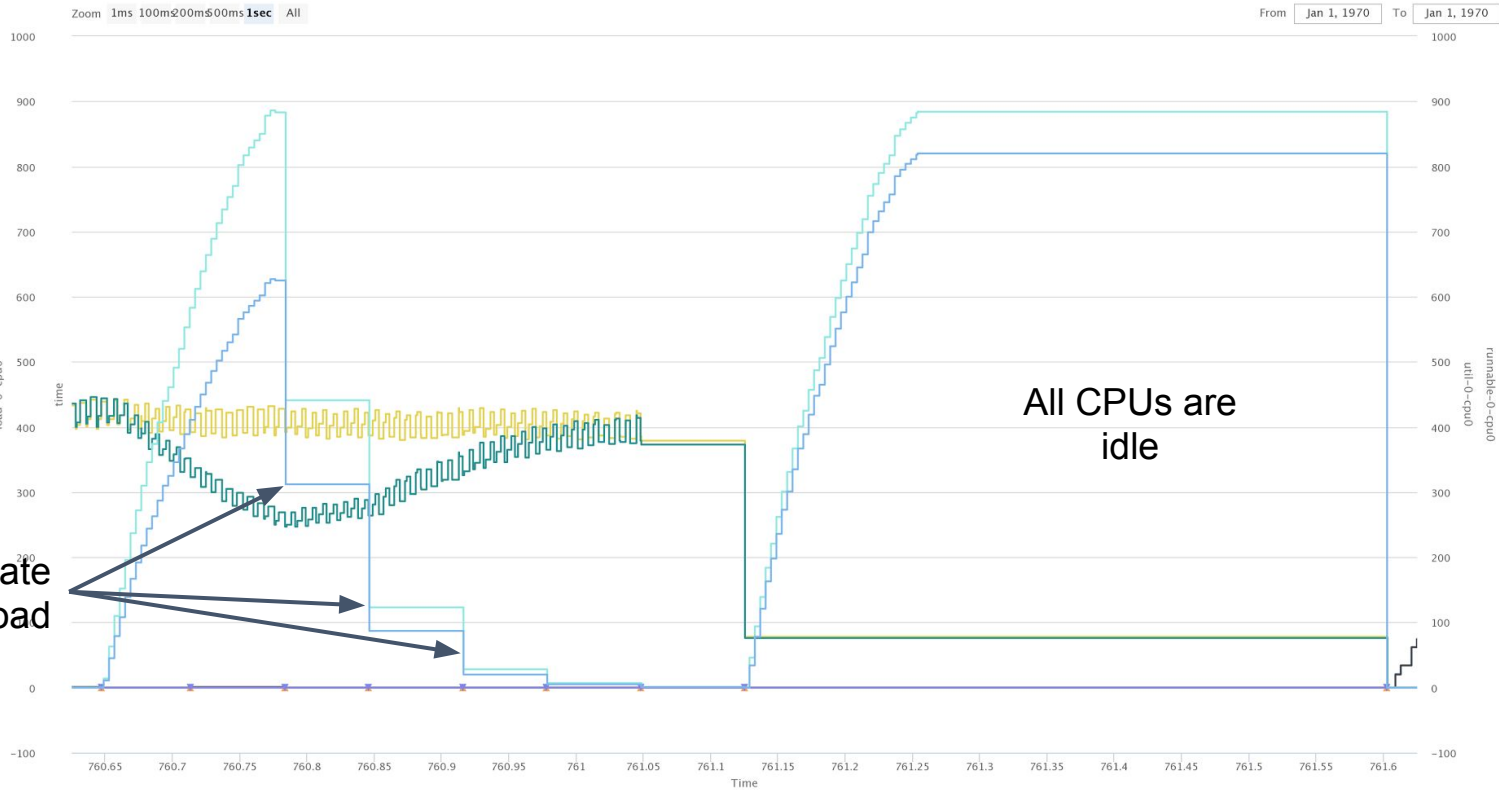Utilization and freq scaling w/ legacy invariance
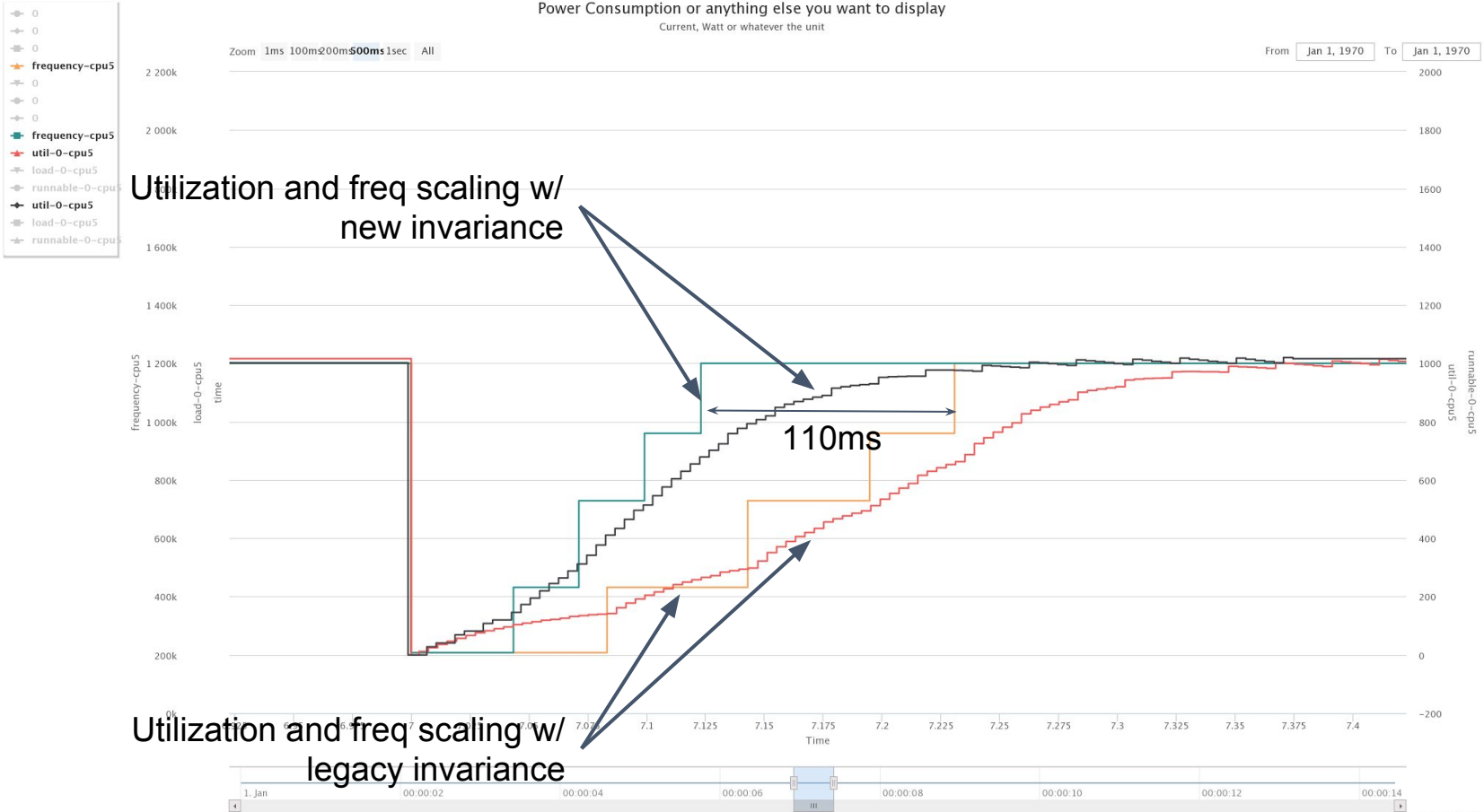
Frequency invariance

- Faster increase
- Same range of variation across freq and micro arch
- Same behavior for Load and utilization

Updating blocked idle load

Power Consumption or anything else you want to display
Current, Watt or whatever the unit

All CPUs are idle

Periodic update of blocked load

Updating blocked idle load

- Periodic load balance updates blocked load periodically
  - Only if load balance is needed
- Need to update block load more frequently
  - Load is used to set share in task group
  - Utilization is used to set OPP (w/ schedutil)
- ILB already update blocked load periodically
  - Kick it more frequently but only to update blocked load
- Kick triggered in select_task_rq_fair
  - If task pinned on 1 CPU, don't kick. Should we ?
  - If RT/DL task don't kick. Should we ?
- If system fully idle, don't wake up
- But not enough
 - When a task wake up, it's already too late

Linaro

LEADING COLLABORATION
IN THE ARM ECOSYSTEM

utilization drop on RT/DL preemption
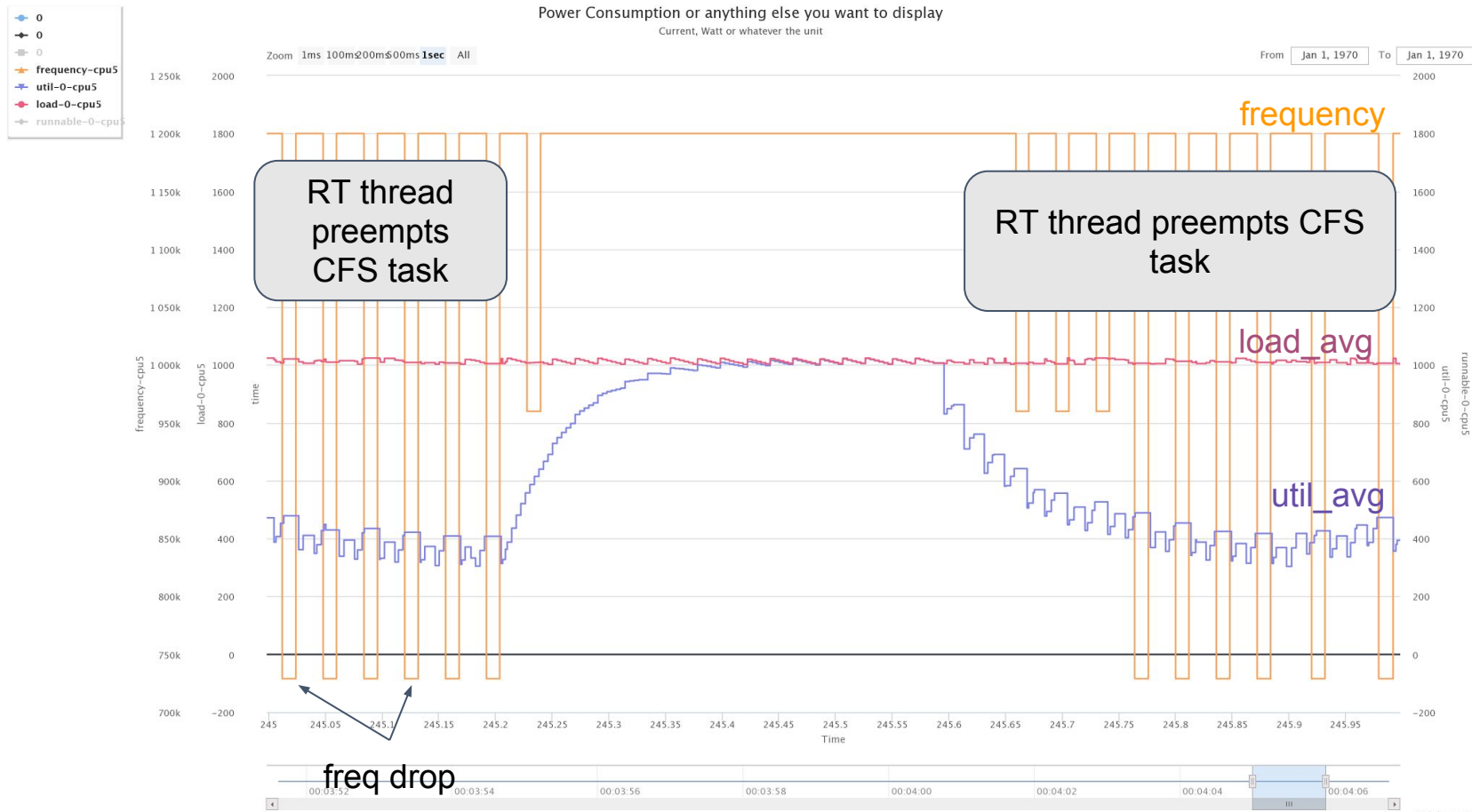
# Prototype RT utilization tracking

- Add a PELT like utilization tracking for rt_rq
  - Only for root rt_rq
- Update in 3 places in RT class
  - pick_next_task_rt
  - put_prev_task_rt
  - task_tick_rt
- And 1 place in CFS
  - update_blocked_averages
- Add CFS+ RT (+ DL) utilization to estimate current utilization
- Must ensure a periodic update of blocked load

utilization drop on RT/DL preemption
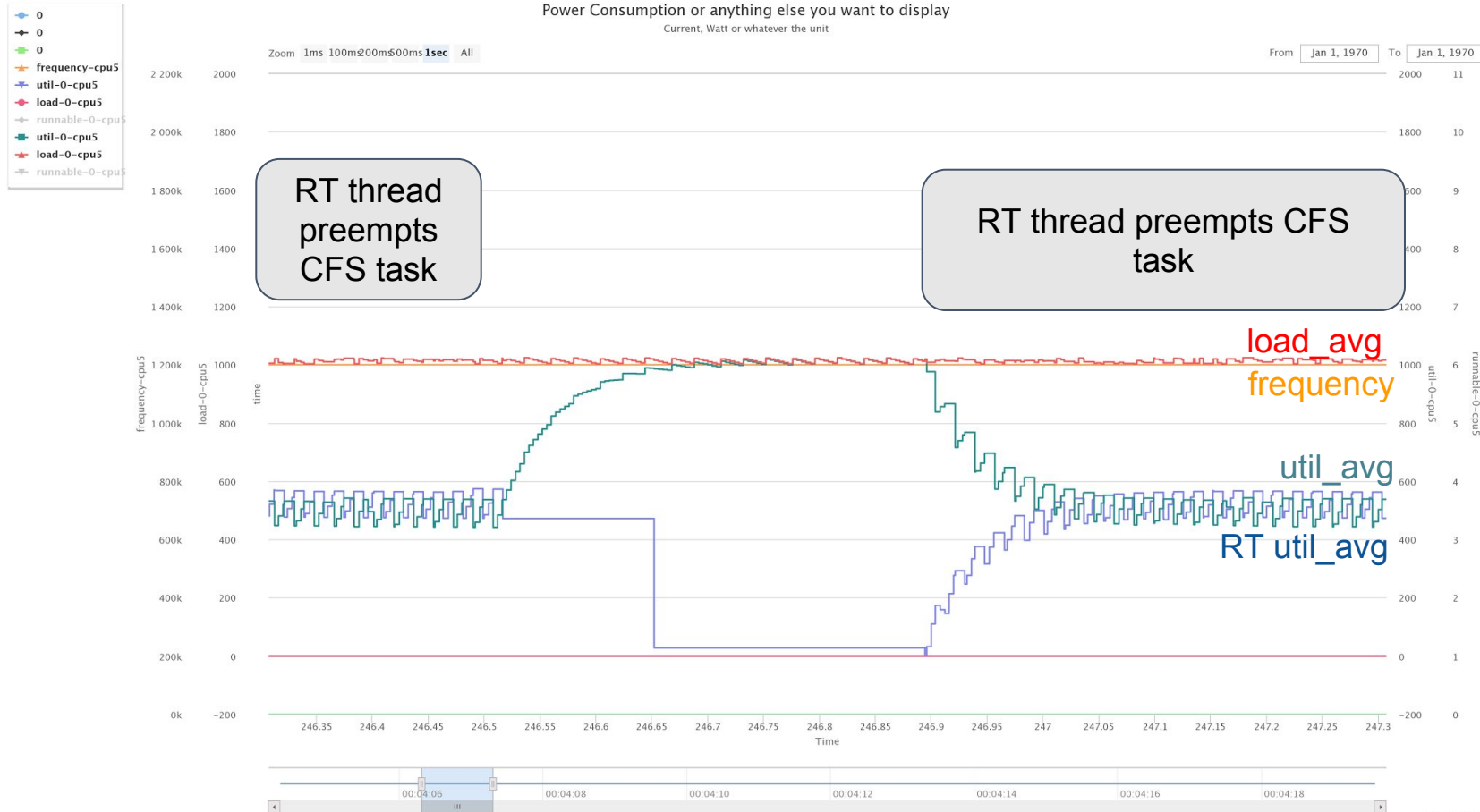
utilization drop on migration
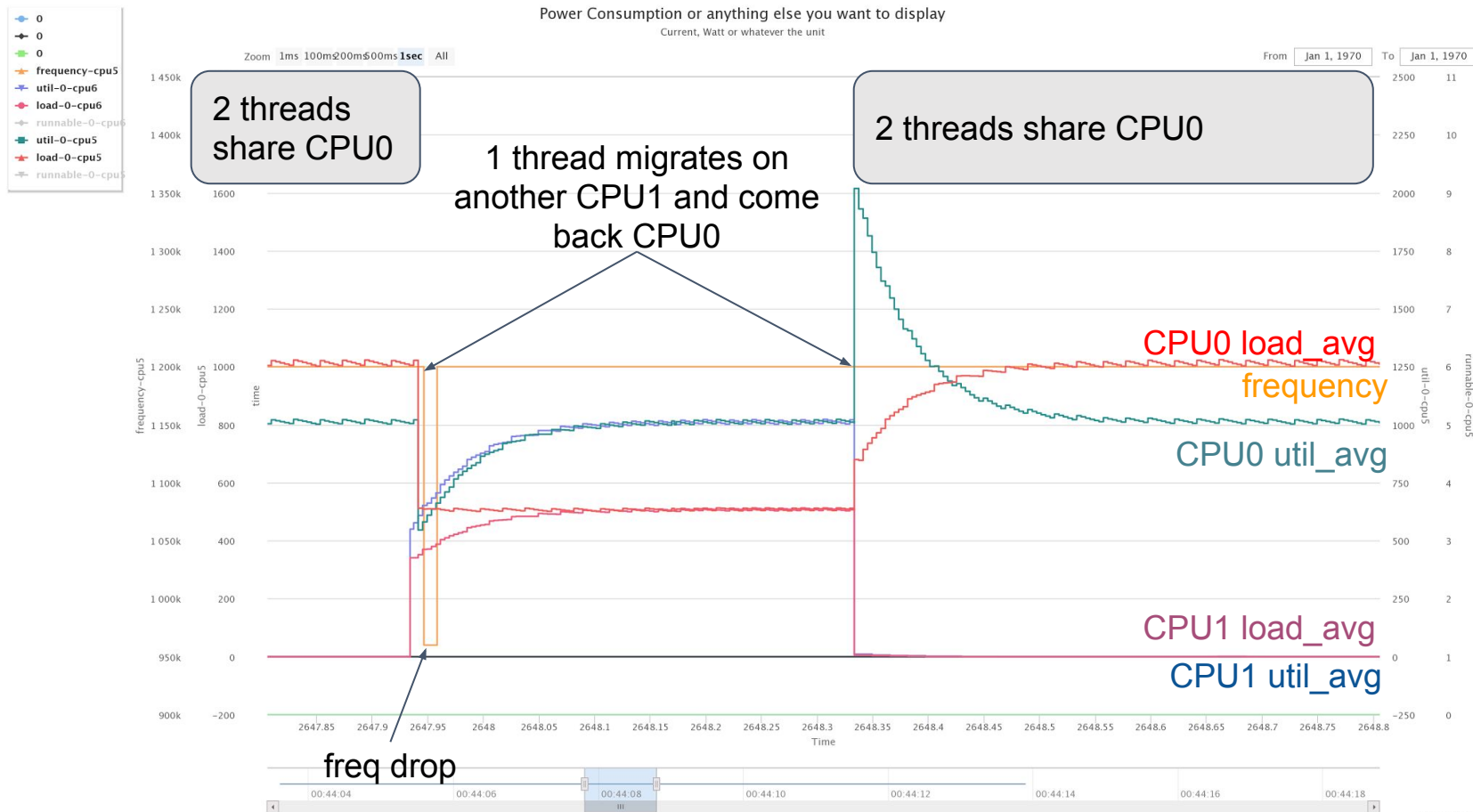
Evaluate waiting time

- How to evaluate that a thread doesn't have all the running time it wants ?

- Not overloaded CPU
  - Can use util_avg without issue
- Overloaded CPU
  - Util_avg is not enough
  - Util_avg doesn't reflect usage of entity
- Use runnable time ?

Others ?

Linaro

Thank You

For further information: www.linaro.org

This is a Standard Content Slide - Enter Title Here

- Enter bullet points here
- Add images on the right or below

Alternative Content Slide

- If you have a more succinct message, use this slide

Linaro

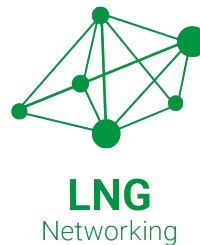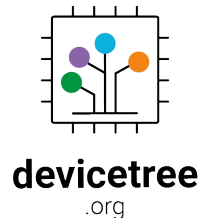LEADING COLLABORATION
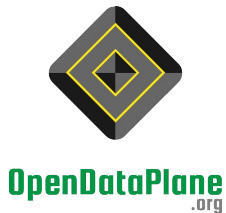IN THE ARM ECOSYSTEM

This Slide is for when Two Columns are Needed

- You can use this for two columns of bullets
- Or you can replace one column with an image or diagram

- This could be a second column of bullets
- Or it could be a table, image or graphic

Use this Slide for a Large Image

Useful Logos

Linaro
connect
Las Vegas 2016

OpenDataPlane
.org

96 Boards

OP-TEE
.org

DeveloperCloud
www.linaro.cloud

devicetree
.org

Linaro

LEG
Server

LHG
Digital Home

LMG
Mobile

LNG
Networking

LITE
IoT/EMBEDDED

Download Hi Res logos from here* to use on your slides

*http://link.linaro.org/logos

Linaro | LEADING COLLABORATION
IN THE ARM ECOSYSTEM

Some Images

- Download Hi Res photos from [here](#)*
  to use on your slides

*http://link.linaro.org/groupphotos

Linaro | LEADING COLLABORATION
IN THE ARM ECOSYSTEM