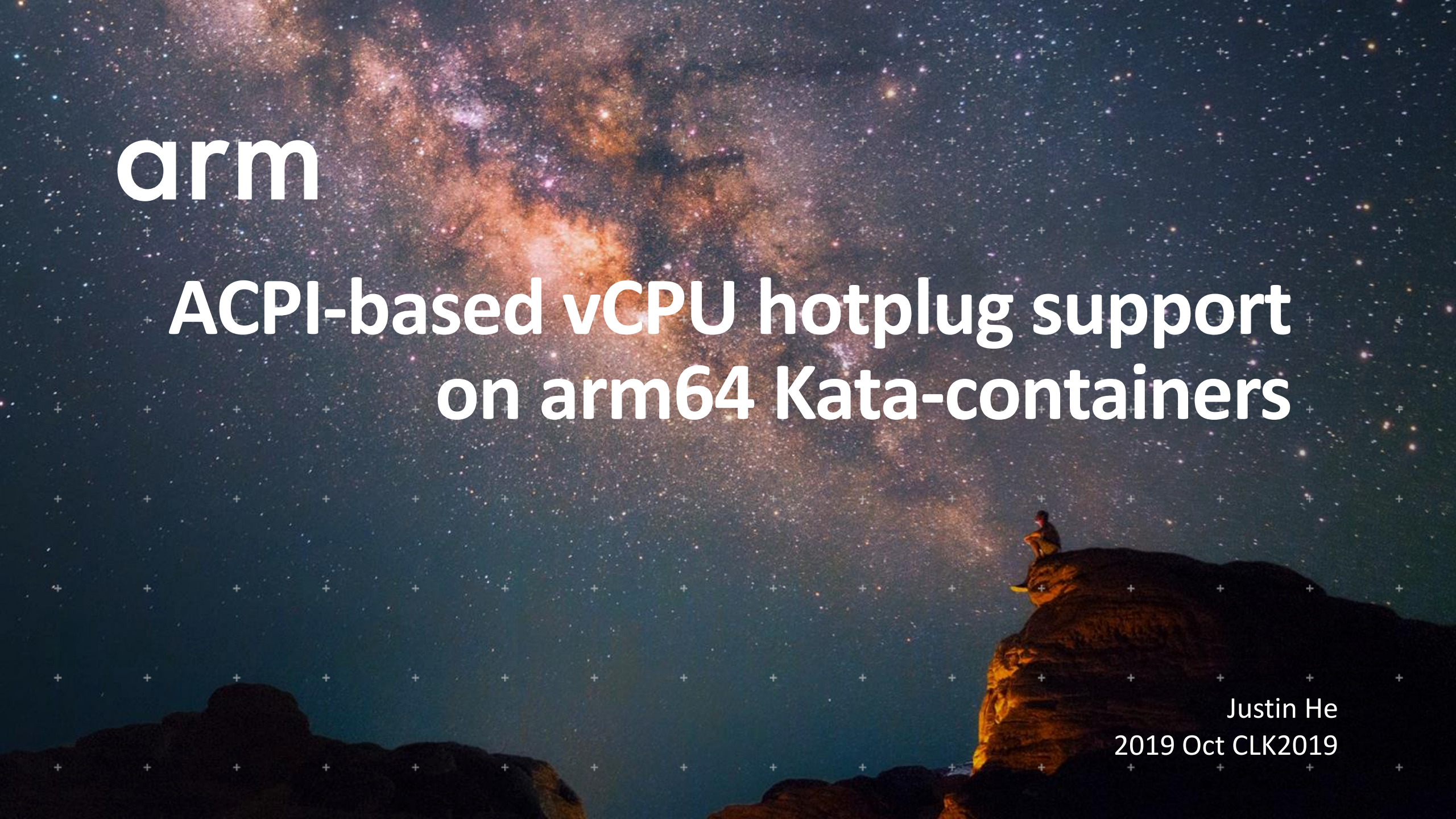


ACPI-based vCPU hotplug support on arm64 Kata-containers

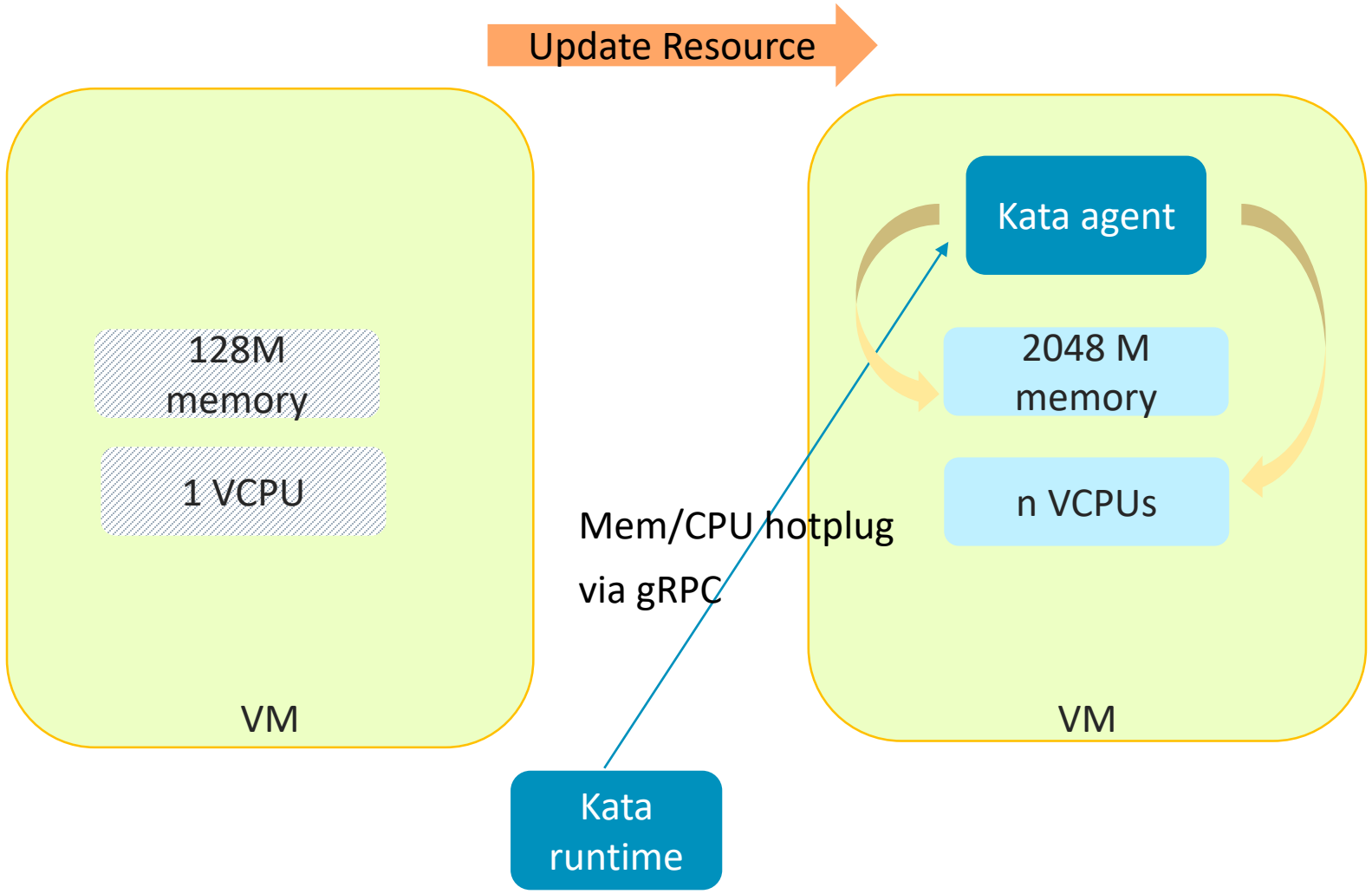
Justin He

2019 Oct CLK2019

Agenda

- The background (What, Why)
- Online/offline solution
- ACPI based solution
- Current status and future plan

Background

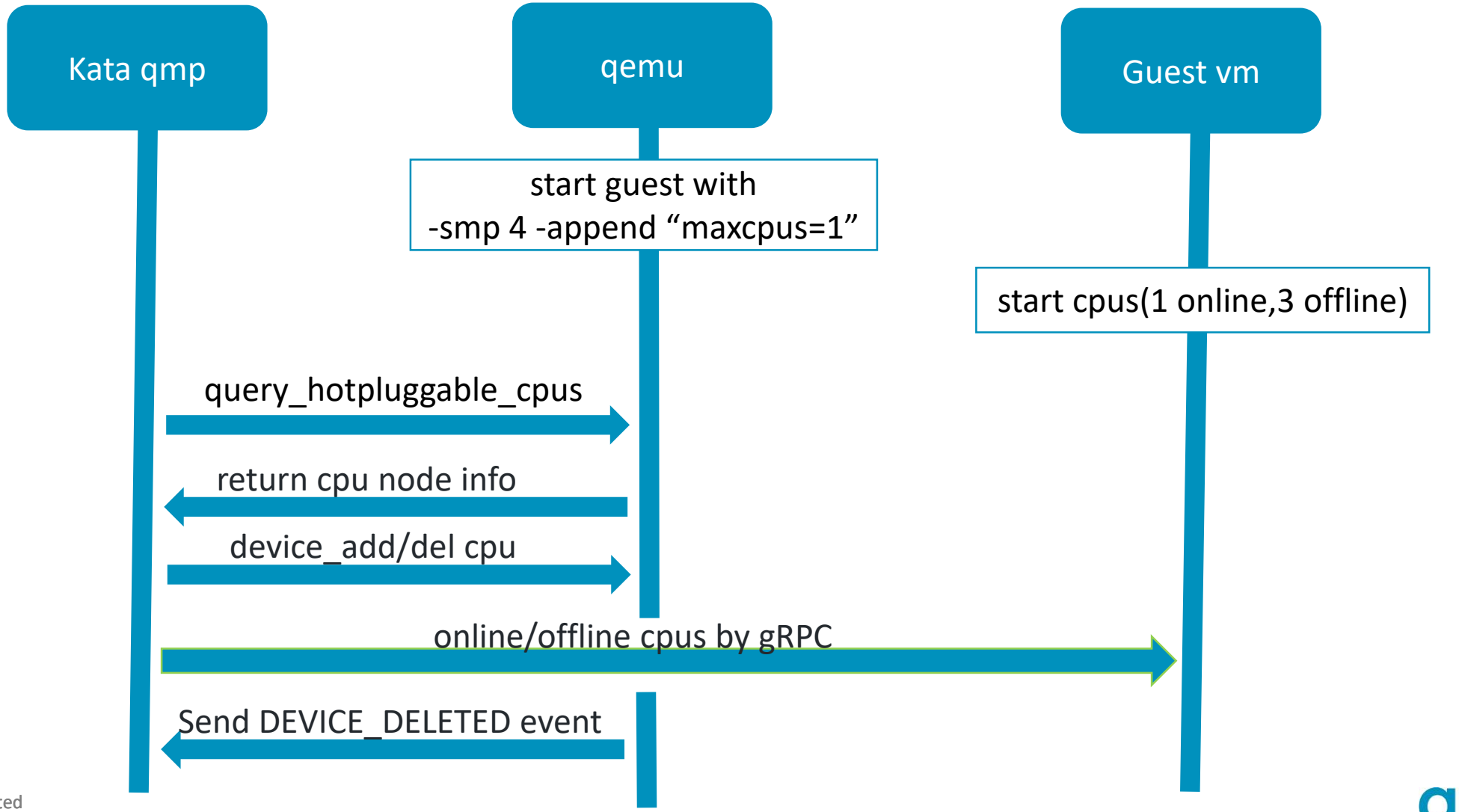


arm

CPU hotplug requirement Kata-containers on arm64

- The target is to support cpu resource update for “docker run” cmd in Kata-containers project
  e.g.
  $ docker run --runtime=kata-runtime -ti --cpus 2 debian bash
  $ docker update --cpus 4 foo
  $ docker update --cpus 2 foo
- Mentioned by Mark Rutland & Christoffer Dall (commit 716139df):
  - Each CPU has a notional point to point link to the interrupt controller (to the redistributor, to be precise), and this entity must pre-exist.
  - When the vgic initializes its internal state it does so based on the number of VCPUs available at the time.

Online/offline solution



Disadvantages:

- the docker container density on arm64.
- an orchestration workaround that neither the hypervisor nor the architecture support.
- security drawback for the containers

ACPI-based vCPU hotplug support on arm64

- Protocol depends on platform
  - ACPI (x86 & ARM)
  - PAPR events (POWER)

- Some assumptions
  - guest kernel should be started with acpi enabled, fdt vcpu hotplug will not be covered
  - consider hotplug add firstly, then hotplug remove
  - consider TCG mode firstly, then kvm mode (vgic initialization limitation)

Steps to support ACPI-based solution

- Qemu changes
  - add vcpu hotplug infrastructure for arm virt machine type
  - build corresponding ACPI tables compared with X86
  - send ACPI event to notify the new vcpu, GED device is a good choice
  - reserve the MaxCpus vcpu for gic initialization, but the vcpu thread will be dynamically added/removed
- Kernel (guest kernel)
  - change its scanning process for ACPI tables (MADT...) for pluggableCpus, they will not be started up.
  - When users try to request hotplug-add, kernel will recv the ACPI event interrupt, then register a new cpu , start the new cpus by PSCI cpu_on call and add it into scheduler system

Status

- Qemu changes
  - Draft codes are done in TCG/gicv2 mode.
  - vcpu thread can be dynamically allocated.
- Kernel (guest kernel)
  - Huawei (Wang xiongfeng) proposed a RFC series in 2019 June
  - After kernel gets the acpi ged event sent from qemu, it can startup the cpu by PSCI call, handle the cpu registering, and add it to scheduler,

Future plan

- Support kvm/gicv3 mode at qemu side
- Do not conflict the codes with support ACPI cpu hotplug physically on host

arm

Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
תודה