CMS perspective on the evolution of WLCG compute slots

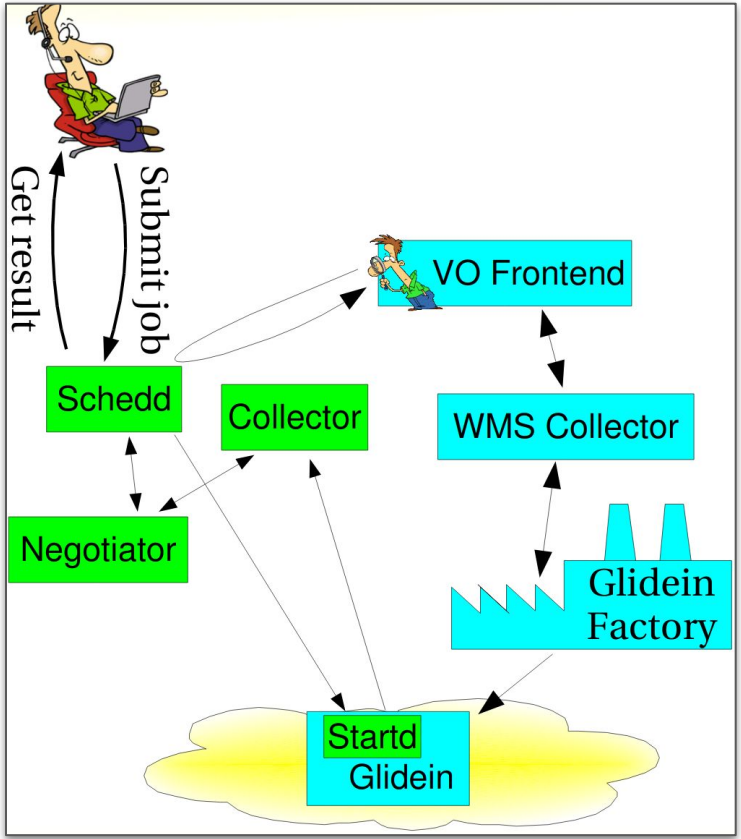
A. Pérez-Calero Yzquierdo for the CMS Collaboration

WLCG Workshop, DESY, 15th May 2024

Outline of the talk

- How CMS requests and uses compute slots from the Grid
- CMS perspective on the potential evolution of Grid slots: high-memory slots, larger slots, whole-node allocation.

Resource allocation for CMS: dynamic HTCondor pools



<u>Late binding pilot-based model</u> with a single type of pilot jobs for all resources and all workloads

- CMS pilots are multicore
- CMS pilots manage multiple types of workloads (type, users, resources…) simultaneously

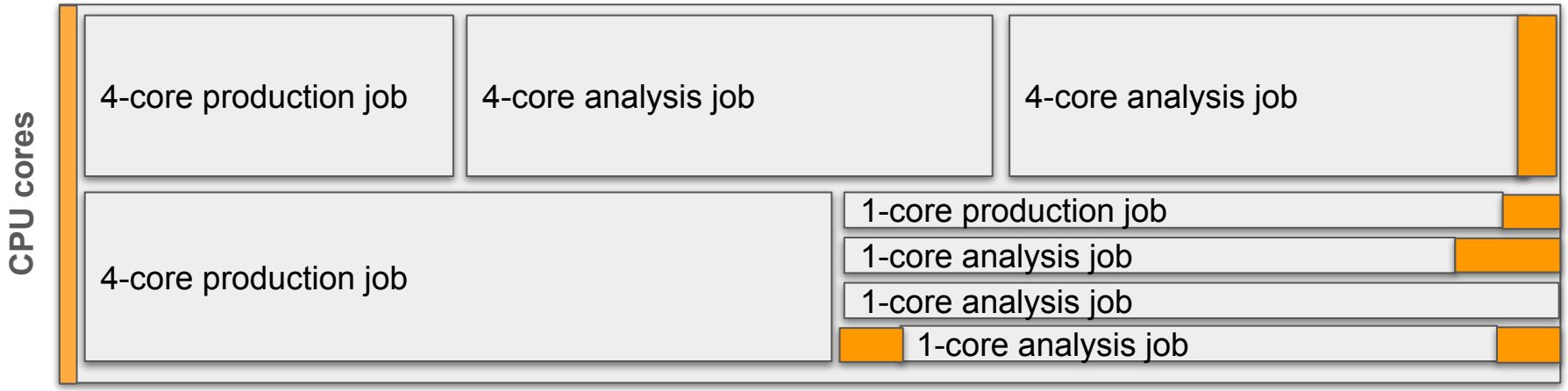
Resource allocation and use based on two **matchmaking** stages:

1. Acquire resources based on GlideinWMS submission of pilot jobs to compatible Grid CEs

2. HTCondor matchmaking of payload jobs to compatible slots

Multicore pilot model in CMS SI

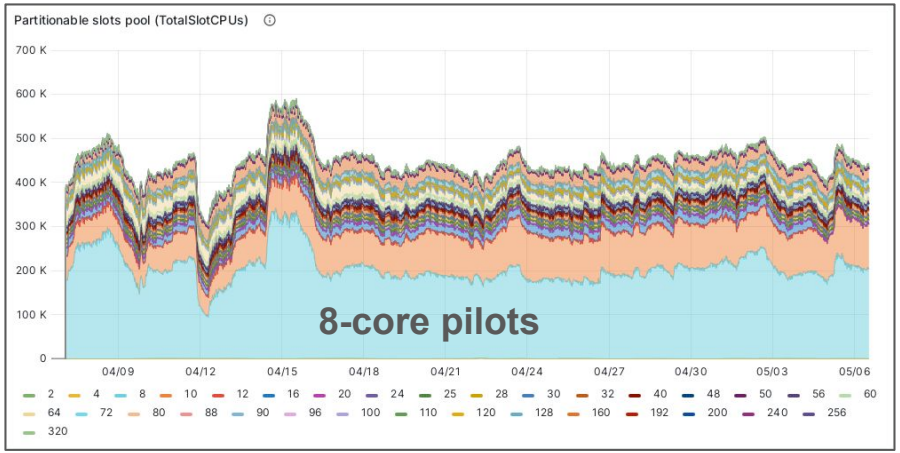
- HTCondor **partitionable slots** allow CMS to execute multiple payload jobs concurrently and consecutively for the duration of the pilot lifetime (typically 48h).
- This model was adopted for the LHC Run 2 and **expanded and refined** since,
  - E.g. to improve the scheduling efficiency within pilots
- Scheduling of individual payload jobs into the resource slots is managed by CMS, not the sites:
  - **Flexibility of the model to better support CMS priorities**
  - **…but any scheduling inefficiencies are "charged" to CMS**



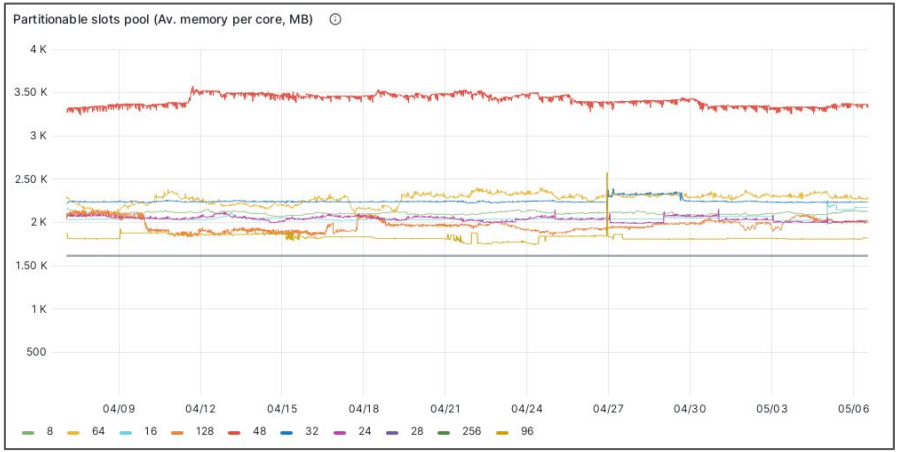
The CMS Multicore pilot model in action (I)

Acquiring resources:

- CMS mainly acquires CPU via **pilots on 8-core slots from WLCG sites**
- CMS model **flexibility** to **accept and use other core-counts larger than standard 8-core (10, 16, 24…)**
- This already includes **whole-node slots** (from some **WLCG** sites and also from **HPC** facilities)
- **Memory/core** in the slots CMS acquires is generally **close to the nominal 2 GB/core**:
  - CMS has **no request** for special high-mem slots (but we can use them)



Number of CPU cores per pilot



Av. memory/core per pilot core size

The CMS Multicore pilot model in action (II)

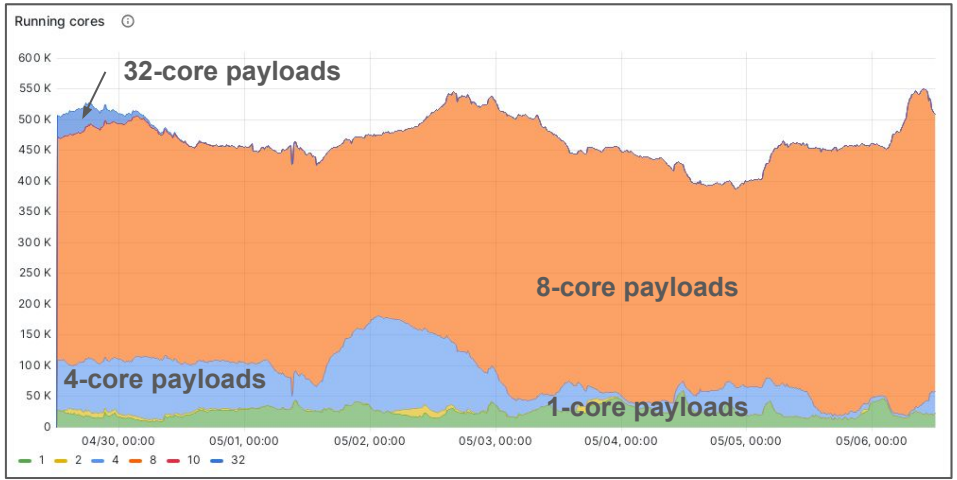
Using resources: Pilots are fragmented into dynamic slots matching job resource requests (CPU, memory, etc)

- **Core-count diversity**: mainly multicore jobs (**4-core, 8-core)**, with some larger requests (e.g. recently **32-core** jobs)
- **Memory per core well adjusted to the 2 GB/core reference value**

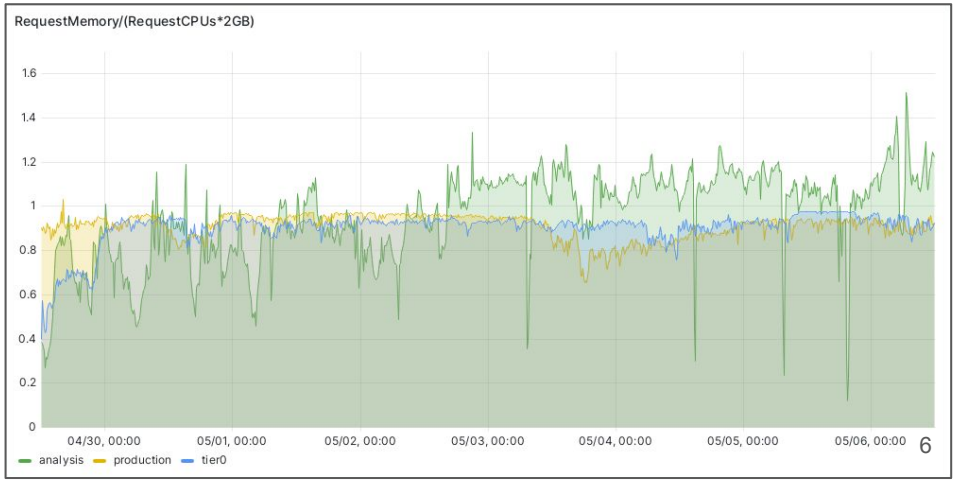
Some thoughts:

- **Does CMS require high-memory slots? Not really**: jobs are well adjusted to the 2 GB/core reference, and the exceptions are taken care by the internal partitioning of the multicore pilots
- **Can CMS benefit from slots with >8 cores? Yes,** our HTCondor pool can integrate them and use them, some payload jobs are already doing multicore >8

Payload jobs running by CPU cores



Payload jobs running by memory/core wrt 2 GB



Whole-node scheduling in the CMS model

CMS is already using whole-node slots from a number of sites, mainly exclusive clusters to CMS (in the US Tier-1 and Tier-2 sites) and from HPC facilities

- Main advantage: Bigger slots represent **increased flexibility for CMS pilot model** on how to dynamically partition resources according to the payload jobs needs
  - **Help CMS getting unusual requests done**!
    - Can accommodate **jobs requesting more than 8-cores** (e.g. simulation with very low gen efficiency producing reasonably sized files while keeping job execution time under control)

- Caveat: **Internal draining** at the end of pilot lifetime may result too wasteful for whole-node slots if the max allowed runtime is kept at 48h
    - To keep efficiency high, whole-node slots **lifetime should preferably be extended from 48h to several days**

Summary

- CMS **model** is based on a single type of **multicore pilot**, capable of handling all types of jobs and partition resources dynamically.
  - **Great flexibility**
  - **Inefficiencies** in internal scheduling **assigned to the VO**

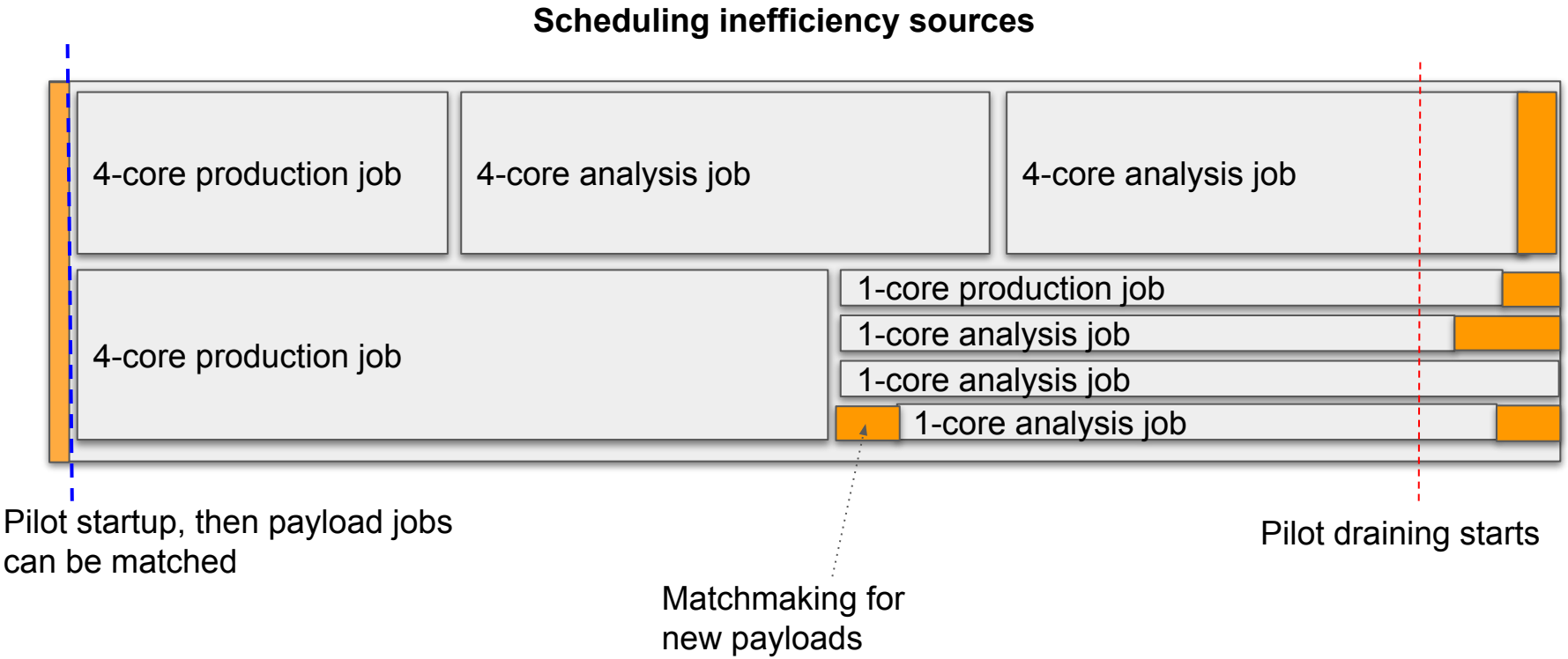
- Does CMS **require high-mem slots**? **NO**
  - Our **multicore jobs** in general **do not require more than 2 GB/core**
  - If exceptionally needed, multicore pilots can mix diverse payload types and provide higher than 2 GB/core slots in regular resources, for a small inefficiency hit

- Can CMS **make use and benefit from 16-core slots? YES**
  - Bigger slot, more flexibility for our model
  - We are already accessing and using slots larger than standard 8 cores
  - Some of our payload jobs already need more than 8 cores

- Can CMS **make use and benefit from whole-node slots? YES**
  - We are already using them at several sites
  - Would preferably get them allocated for N days to minimize impact of the draining phase on resource utilization efficiency

Backup slides

Scheduling efficiency in CMS pilots (I)



Scheduling efficiency in CMS pilots (II)

Slot utilization efficiency for CERN and Tier-1 resources over the last 7 days: typically ~95%