Attempting to measure the Graviton

Benchmarking is hard

- Modern CPUs are complex
  - Cache
  - Dynamic frequency scaling
  - NUMA (resource grouping)
- Operating System interference
  - Interrupts
    - Disk I/O
    - Network packets
- Background processes

Benchmarking in the cloud is harder

linaro Connect

- Virtualization and containerization overhead
  - Hypervisor also uses CPU cycles
- Performance counters are not always available on Virtual Machines
- Shared resources and noisy neighbors
  - CPU cache pollution
  - Memory bandwidth contention

Is it possible to get consistent results in the cloud?

Instance characteristics

- No turbo boost
- (At least) 2GB of RAM per core
- 1 NUMA node
- Private L1/L2 cache
- Shared L3 cache
- SSD/SSD-based (e.g., Amazon EBS) disk

Setup

- SPEC CPU® 2017
  - Stresses CPU, memory and compiler
  - 3 iterations, to measure variance
  - SPECrate® 2017
    - 1 copy, CPU affinity set
  - SPECspeed® 2017
    - All cores used

- Ubuntu 24.04 LTS
  - Docker container

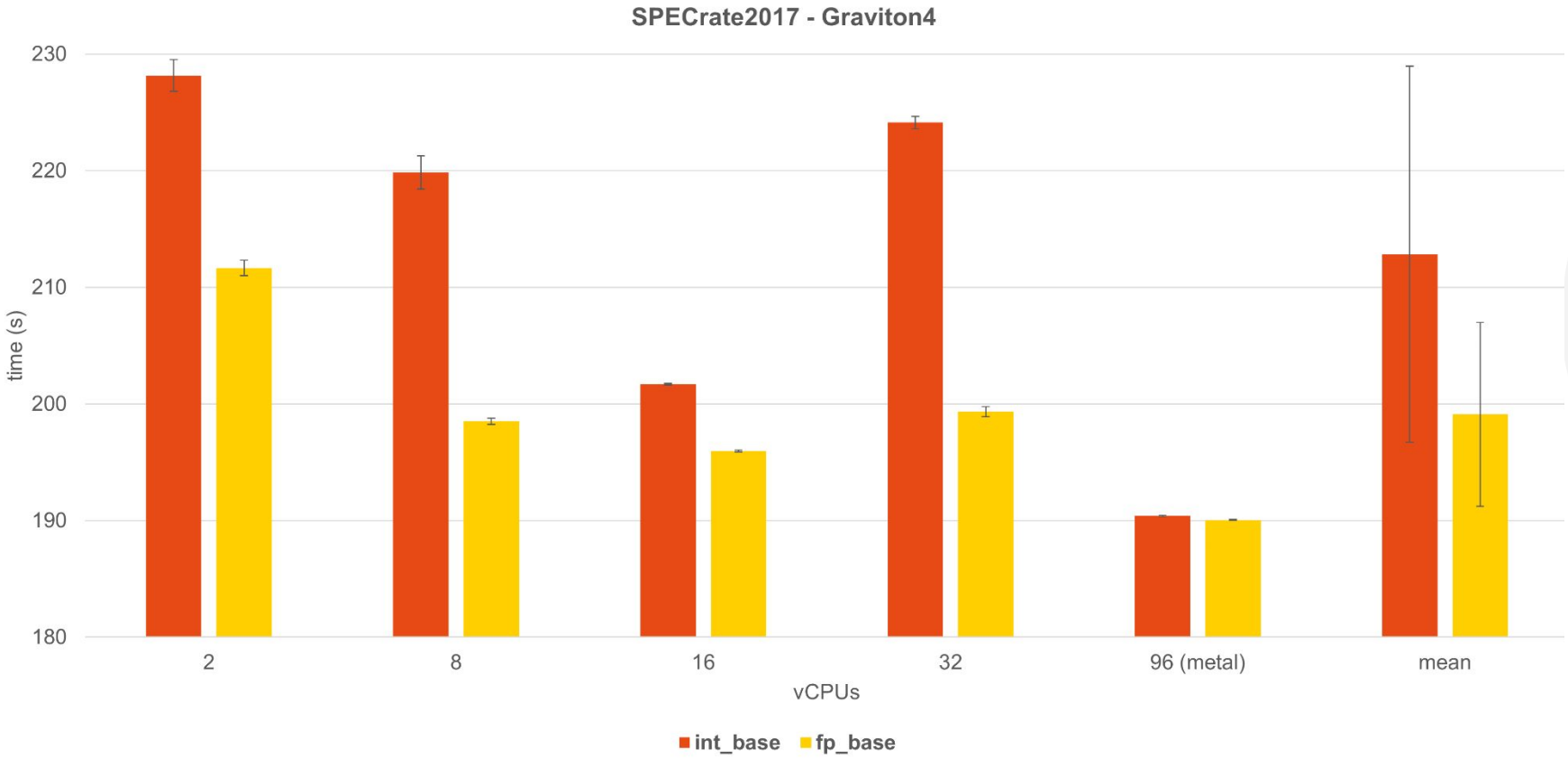
- LLVM 20.1.0
  - Flags: -O3 -flto

Result charts

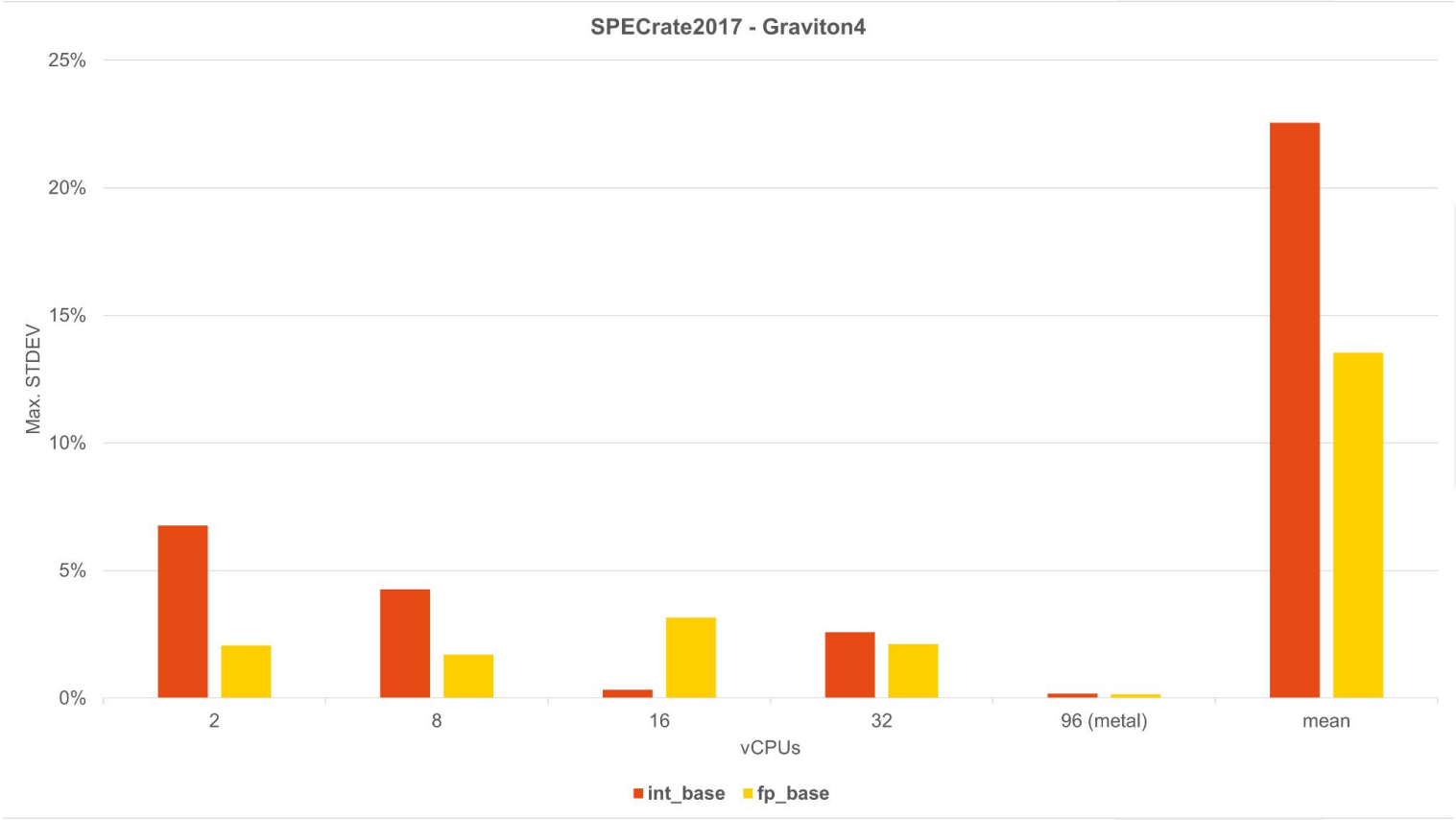
- Instance times and standard deviations are the geometric means of all tests
  - Ex: SPECint time = geomean(500.perlbench_r, 502.gcc_r, ...)

- Mean times and standard deviations are calculated over instance times
  - Ex: SPECint mean time = mean(instance1, instance2, ...)

AWS - Graviton4



AWS - Graviton4 - worst cases



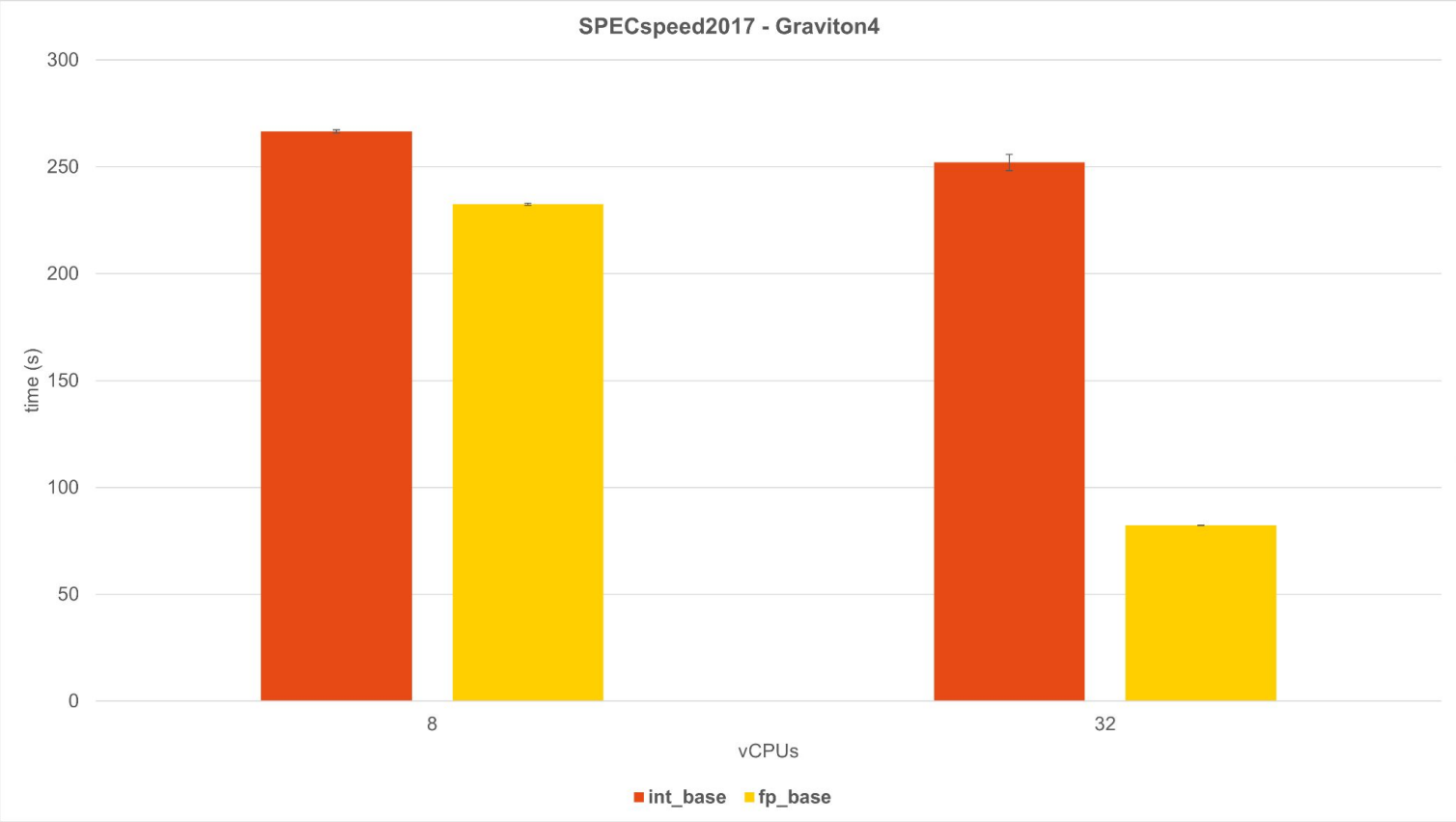
AWS - Graviton4

- Worst case
  - Individual benchmark with the largest standard deviation in a given instance
  - Mean: benchmark with the largest standard deviation across all instances

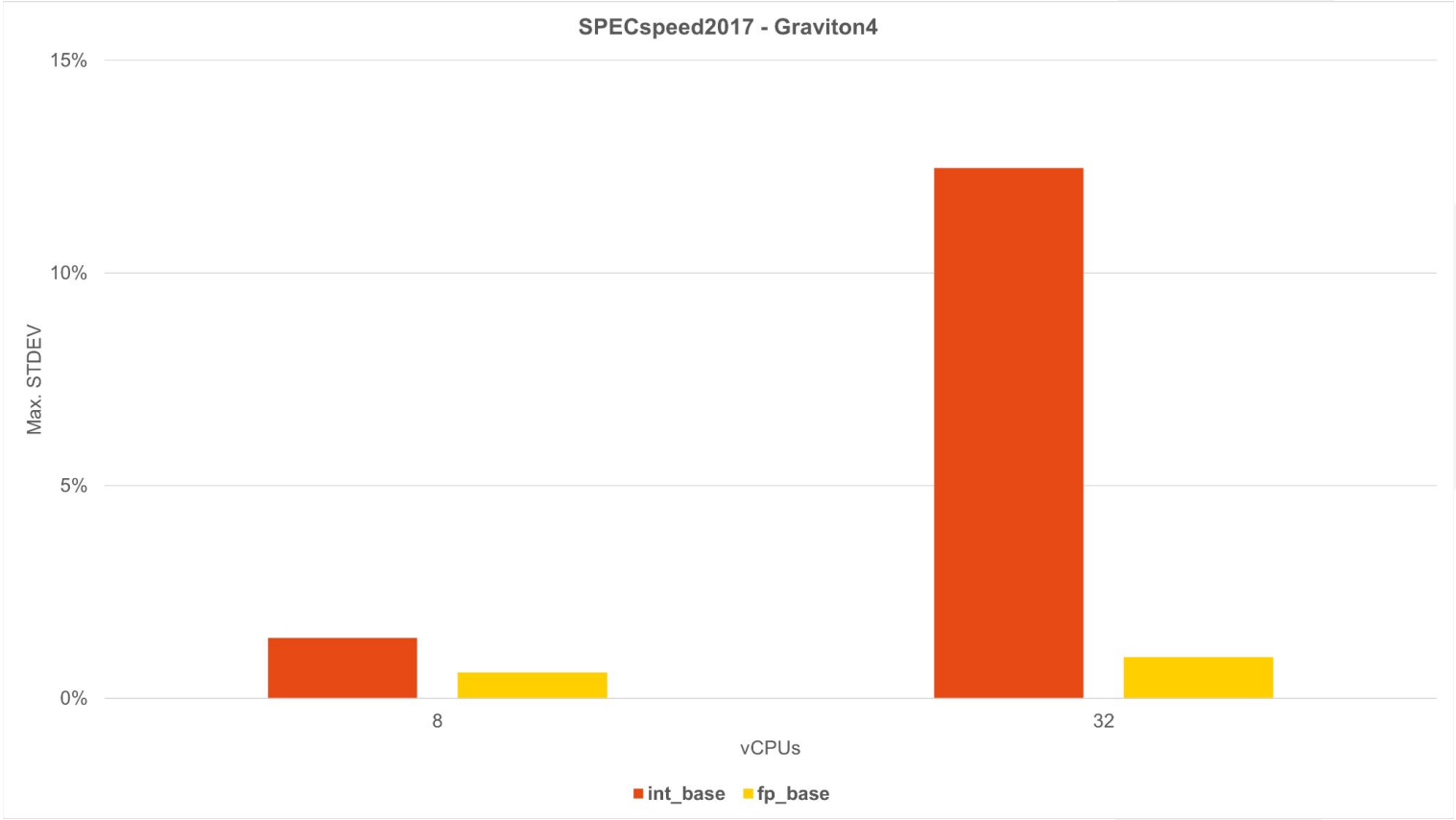
- Bare-metal dedicated instance (baseline)
  - No interference
    - Physical hardware doesn't run instances from other AWS accounts
  - No hypervisor
    - OS runs directly on hardware
    - Still managed by AWS Nitro
  - No Docker container
    - No significant overhead was observed when running inside a container

More details: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/dedicated-instance.html

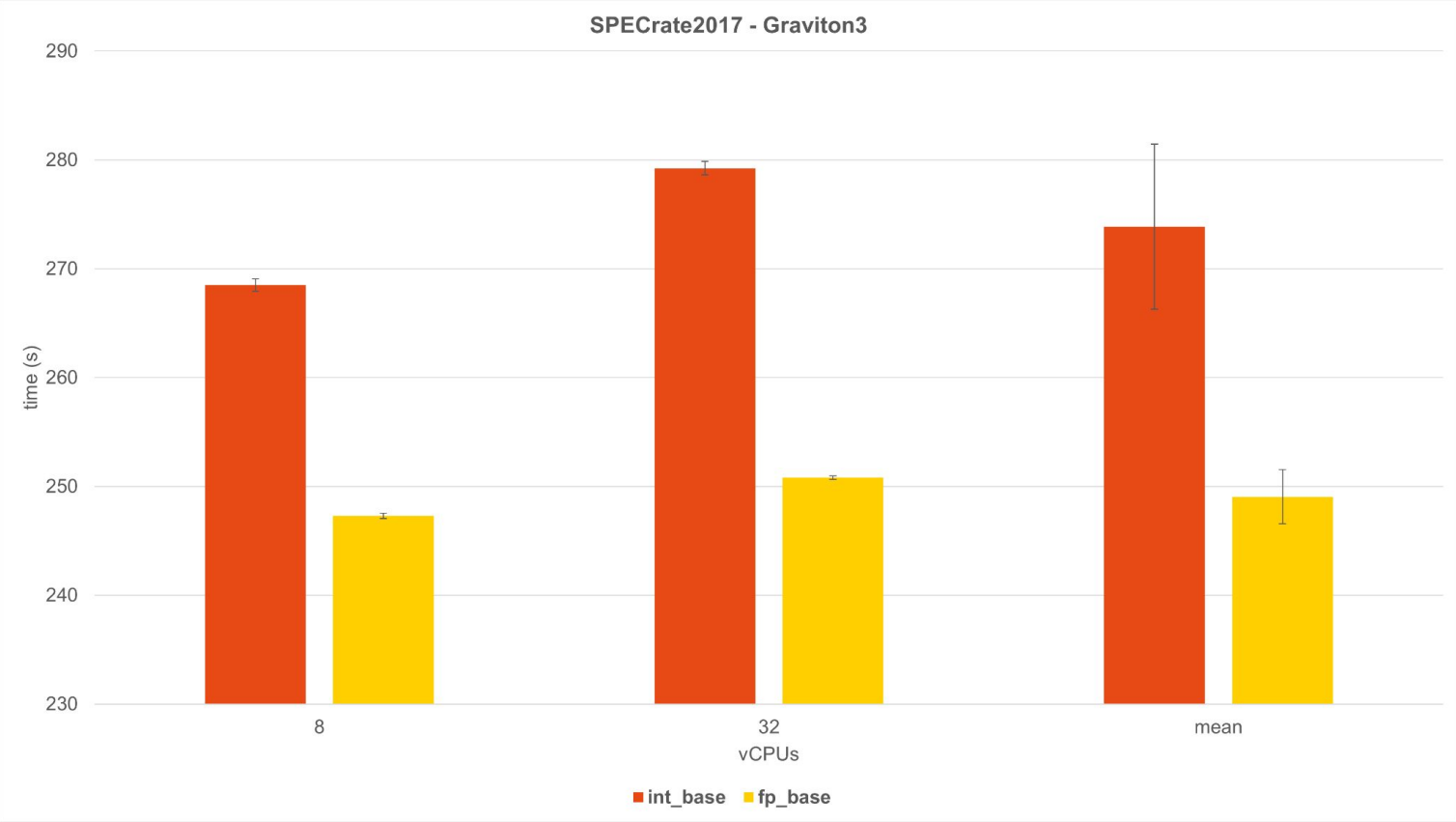
AWS - Graviton4 - SPECspeed®



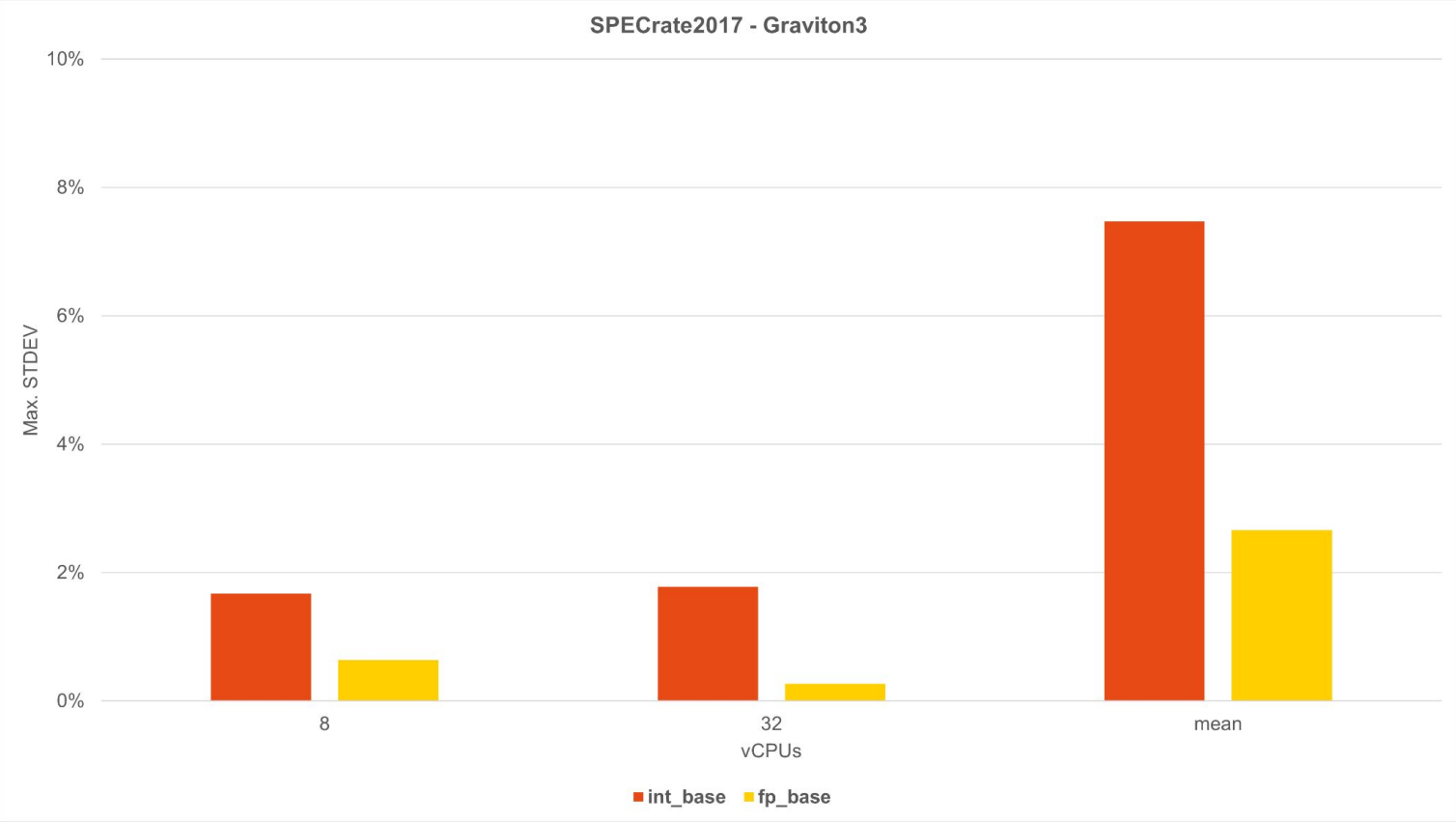
AWS - Graviton4 - SPECspeed® - worst cases



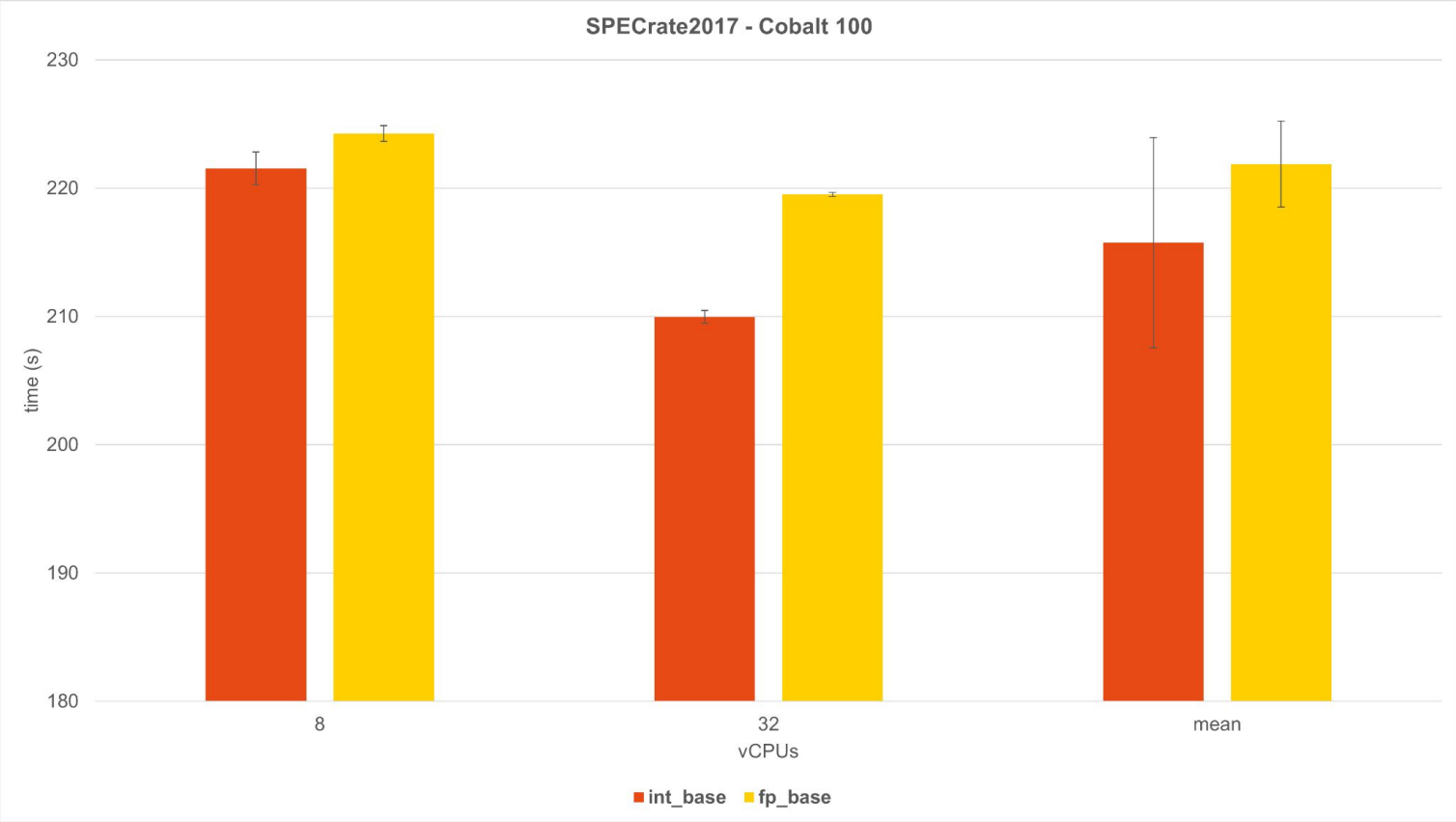
AWS - Graviton3



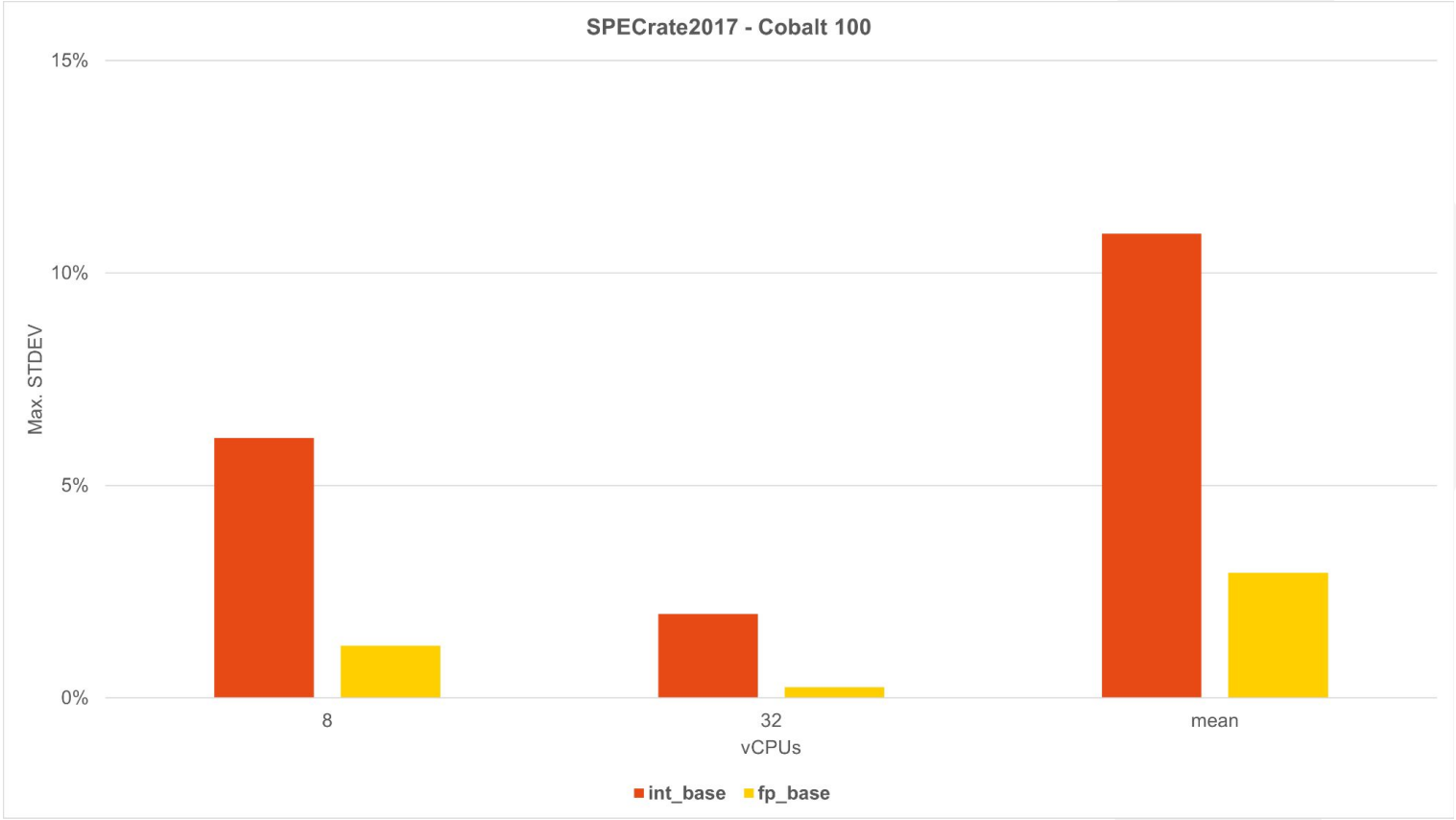
AWS - Graviton3 - worst cases



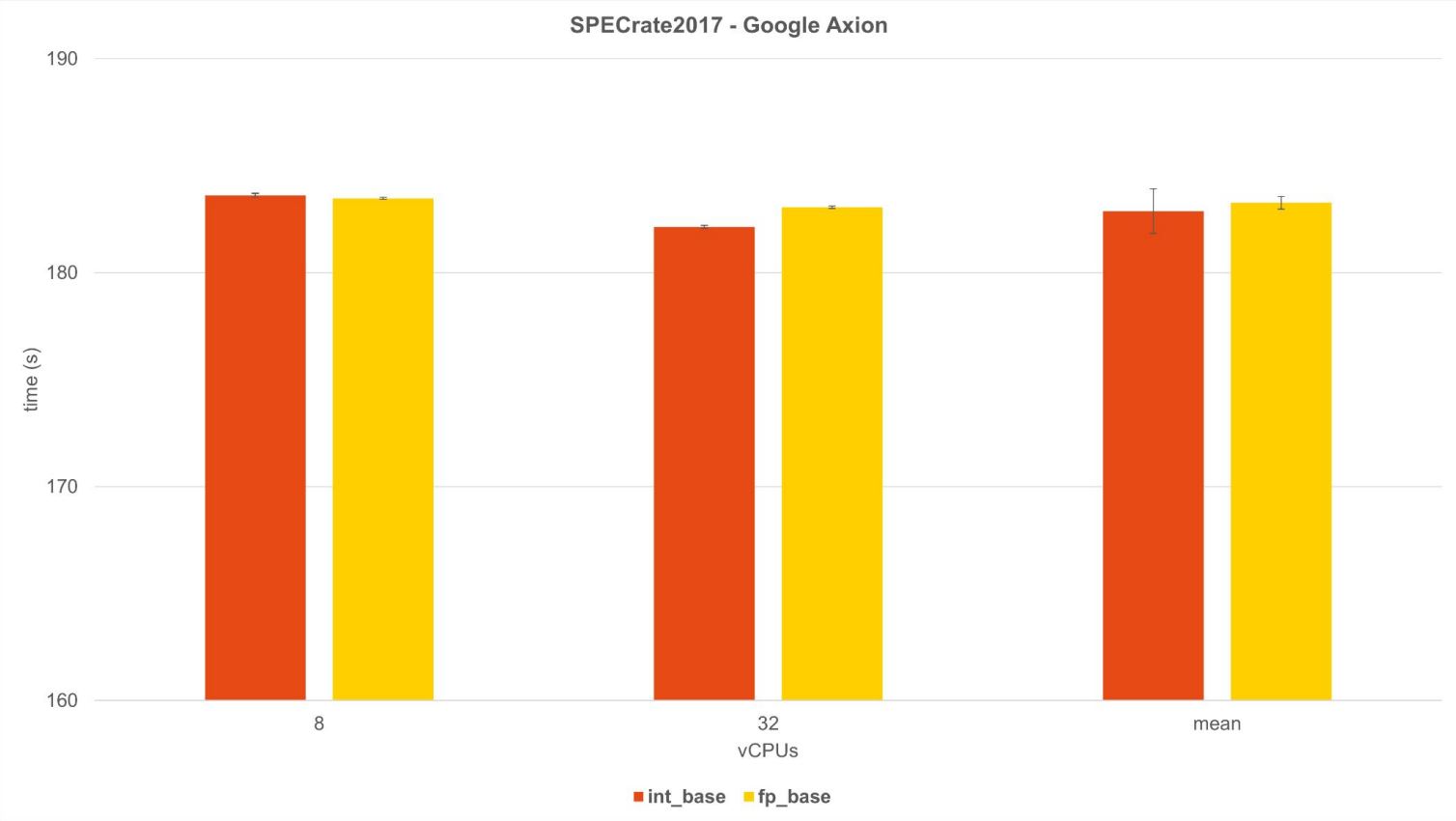
Microsoft Azure - Cobalt 100



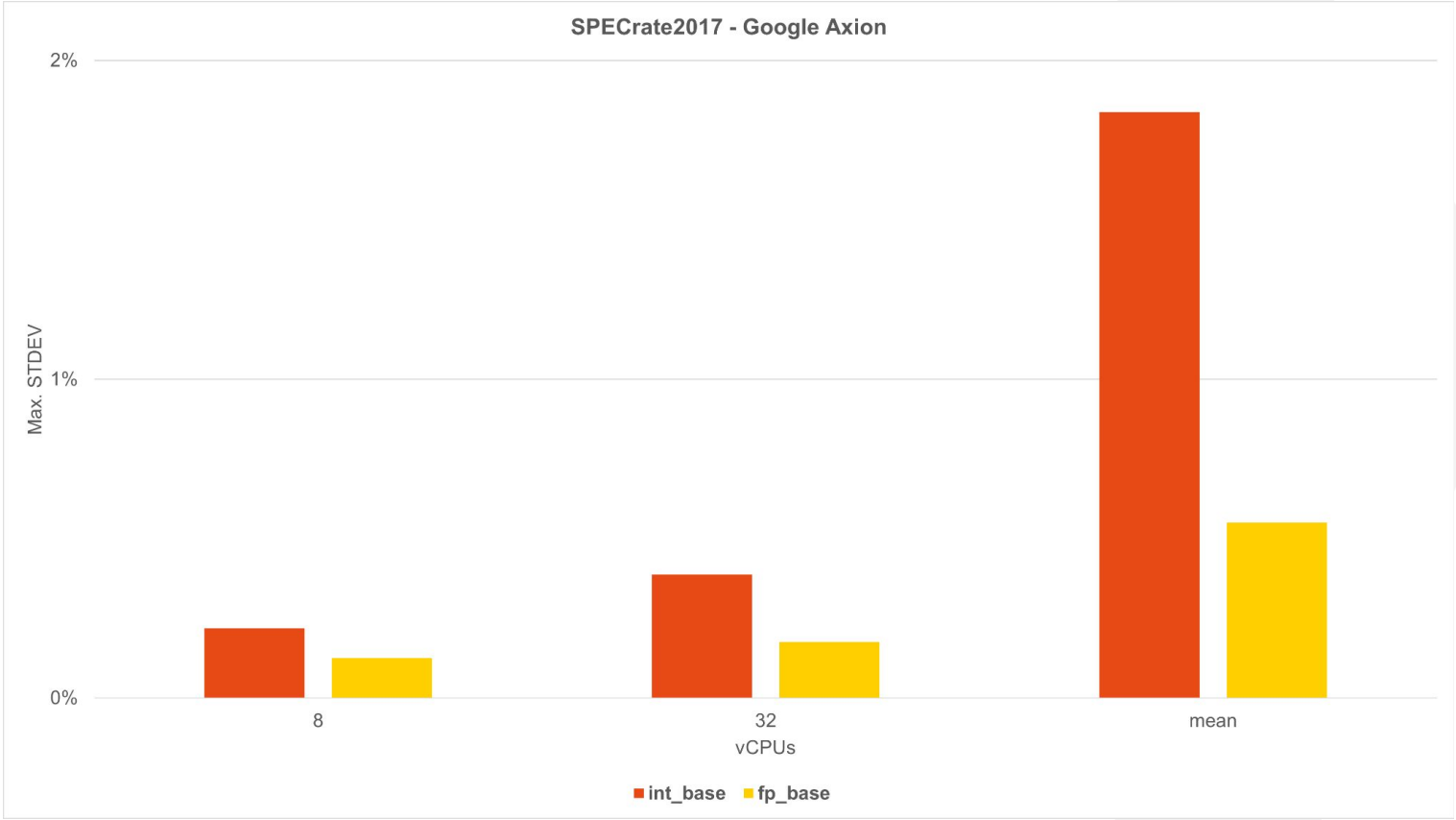
Microsoft Azure - Cobalt 100 - worst cases



Google Cloud - Google Axion



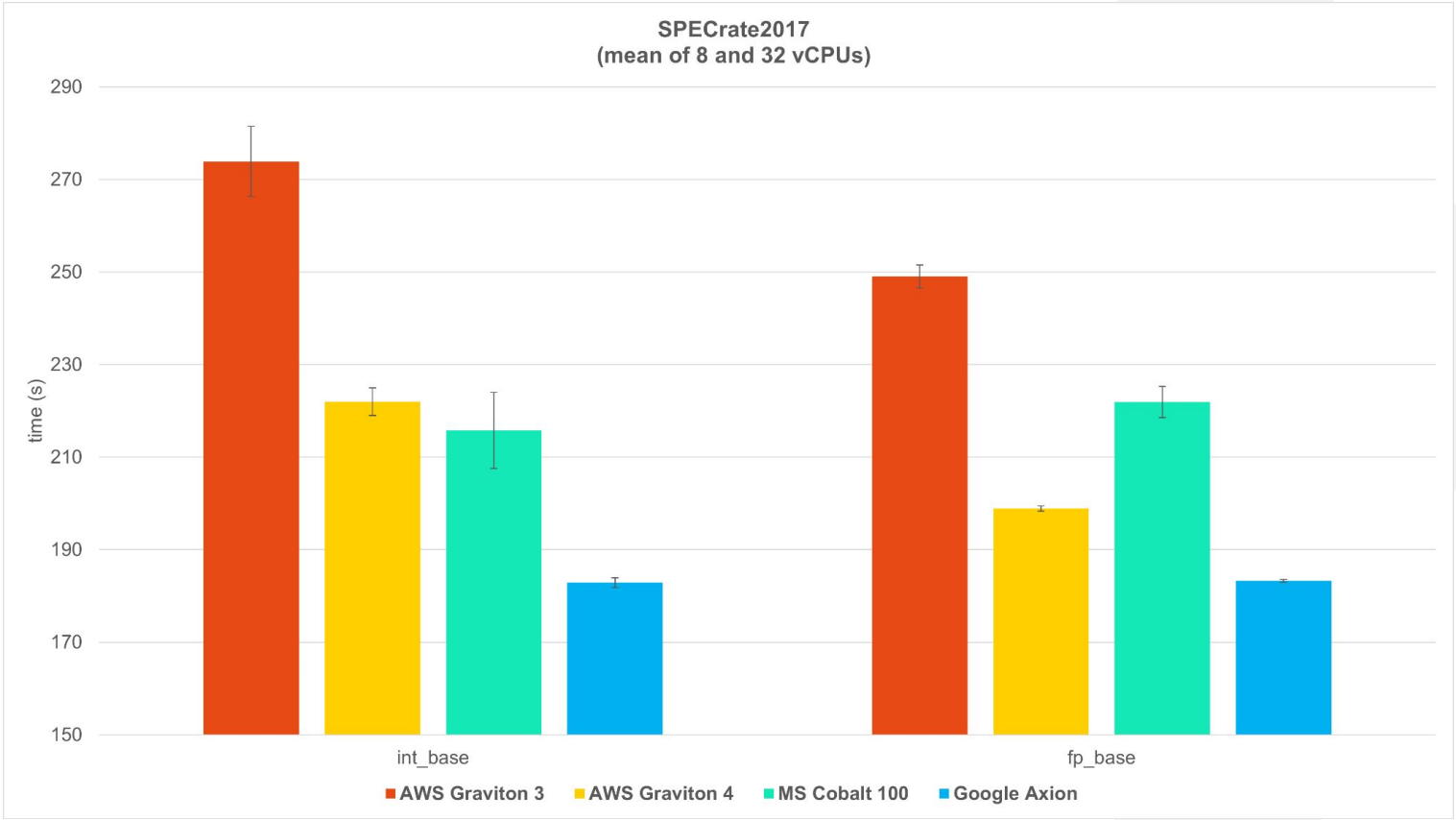
Google Cloud - Google Axion - worst cases



Google Cloud - Google Axion

- Google Titanium
  - Offload network and I/O processing
  - Hardware acceleration for virtualization services

AWS vs Azure vs GCP



AWS vs Azure vs GCP - Notes

- Approximate estimate
- Few runs
- Few instance types
- Subject to random interference from other tenants

Costs

- 32 vCPUs, 2 GB/core:
  - AWS Graviton 3 - 1.16 USD/hr
  - AWS Graviton 4 - 1.28 USD/hr
  - Azure Cobalt 100 - 1.25 USD/hr
  - Google Axion - 1.44 USD/hr
    - Compute-optimized not available, cost for standard (4GB/core)
- 96 vCPUs, 2 GB/core, bare metal dedicated instance:
  - AWS Graviton 4 - 4.21 USD/hr
- Storage and network costs not included

Final considerations

- PMUs worked only on AWS Graviton3/Graviton4

- Variation
  - Usually low in the same instance, in short (few hours) intervals
  - Can be high when comparing different instances or individual benchmarks
  - More vCPUs don't guarantee better performance or lower variation
  - Some workloads are more sensitive than others (e.g., int vs fp)

- Best options for consistent results
  - AWS bare-metal dedicated instances (more expensive)
  - Google Axion instances (no PMU)

Azure dedicated hosts: https://learn.microsoft.com/en-us/azure/virtual-machines/dedicated-hosts
GCP sole-tenant nodes: https://cloud.google.com/compute/docs/nodes/sole-tenant-nodes

Future work

- Check if it's possible to enable PMUs on Cobalt 100/Google Axion
- Test other ARM CPUs
- Check if smaller AWS dedicated instances also have low variation
  - Lower cost
  - Multiple dedicated instances (of the same account) may interfere with each other
- Try to detect performance degradation
- Investigate interference causes and most affected workloads


linaro
Connect
2025
Thank You!