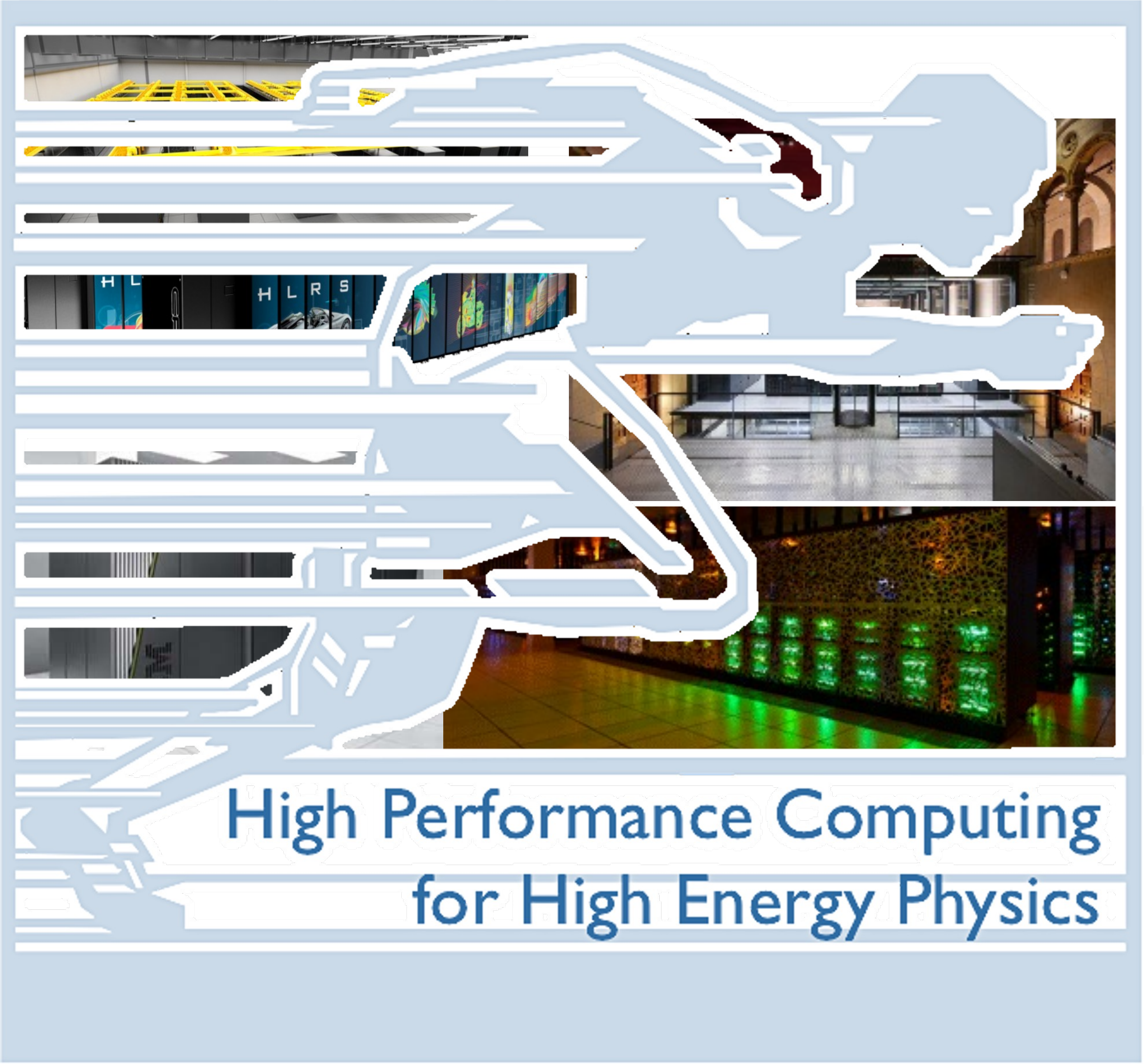
A Deep Dive in The Performance of HepSpec Workflows



Vincenzo Innocente
CERN/SFT

Hepix Workshop, Sept 2022

Machines

- Haswell: (16 cores) Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz
- Broadwell: (24 cores) Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
- Skylake: (32 cores) Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz
- Icelake: (32 cores) Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz

- Haswell went offline: measurements incomplete
- Broadwell clock seems to be locked at 2.45GHz
- Would have been useful to test AMD machines as well
  - Used a workstation for some limited tests

Workflows

- Alice
  - Gen-sim-digi-reco
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/alice-gen-sim-reco-run3-bmk:ci-v0.6-aod -t4
- Atlas
  - Gen (single thread)
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/atlas-gen_sherpa-bmk:v0.2 -t1
  - Sim
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/atlas-sim_mt-bmk:v0.4 -t4
  - Reco
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/atlas-reco_mt-bmk:v0.1 -t4
- CMS
  - gen-sim
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/cms-gen-sim-run3-bmk:v0.6 -t4
  - digi
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/cms-digi-run3-bmk:v0.6 -t4
  - reco
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/cms-reco-run3-bmk:v0.6 -t4
- Juno (single thread)
  - Gen-Sim-Reco
    - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/juno-gen-sim-reco-bmk:v2.0 -t 1
- IGWN
  - singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/igwn-pe-bmk:v0.3 –t4
- LHCb (single threads)
  - gen-sim singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/lhcb-gen-sim-2021-bmk:ci-v0.4 -t1

Methodology

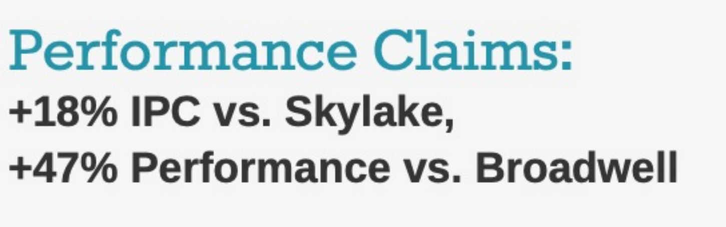
- Use *perf record* / *perf report* to understand WHAT we are actually running and identify hot-spots (single copy)
  - Retuned number of events to reduce impact of initialization
- Use *turbostat* to make a time profile of used resources (full machine)
- Use *perf stat* for a detailed understanding of performance
  - Single copy, full machine with and w/o HyperThread
- Number of events “adjusted” to avoid initialization overhead or too long runs

- All tools run “on metal”: full singularity job profiles
  - perf stat singularity run …

Expectations

- https://www.anandtech.com/show/14514/examining-intels-ice-lake-microarchitecture-and-sunny-cove/3

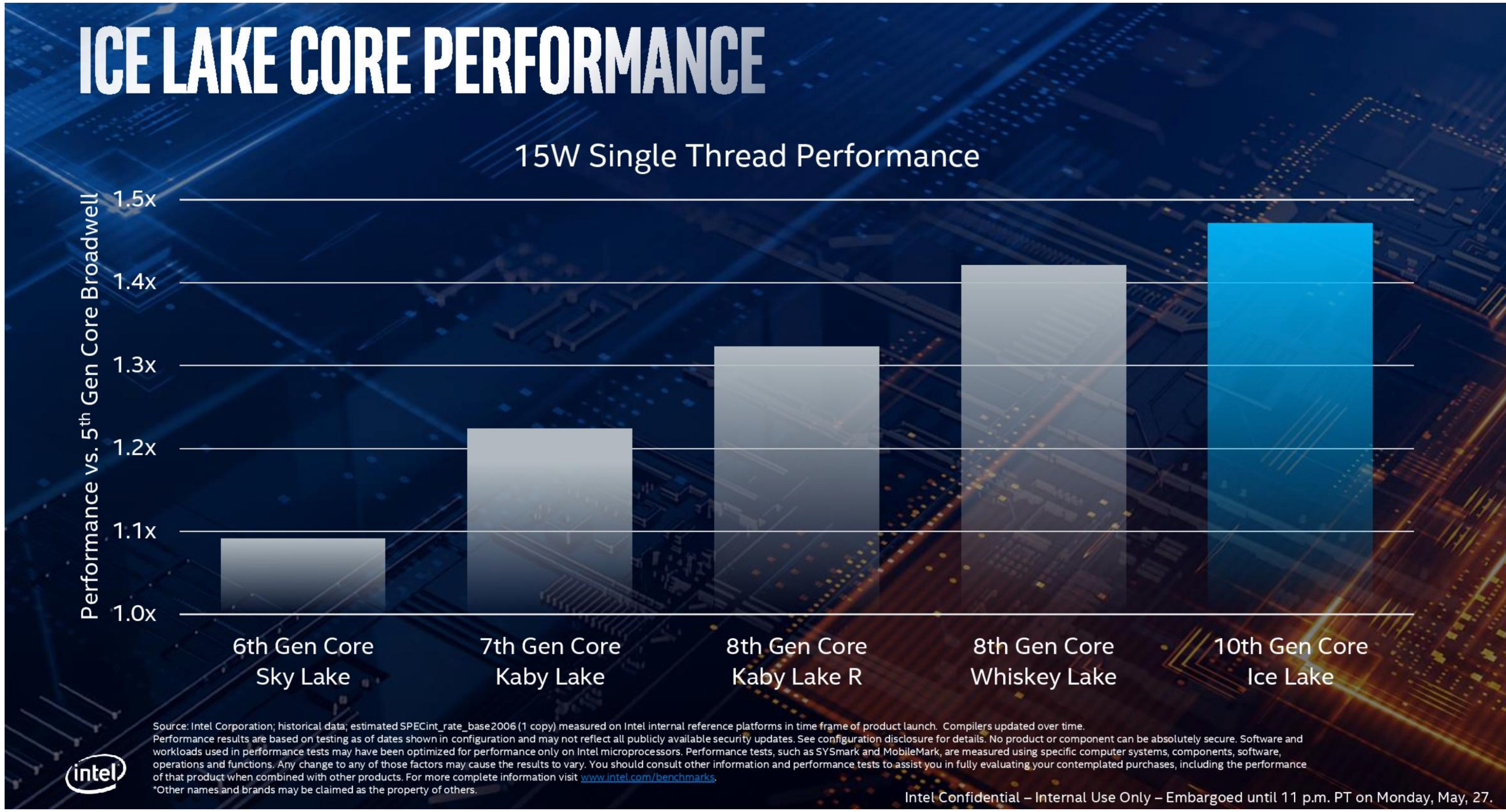
Icelake



Performance Claims:

+18% IPC vs. Skylake,

+47% Performance vs. Broadwell



ICE LAKE CORE PERFORMANCE

Source: Intel Corporation; historical data; estimated SPECint_rate_base2006 (1 copy) measured on Intel internal reference platforms in time frame of product launch. Compilers updated over time. Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

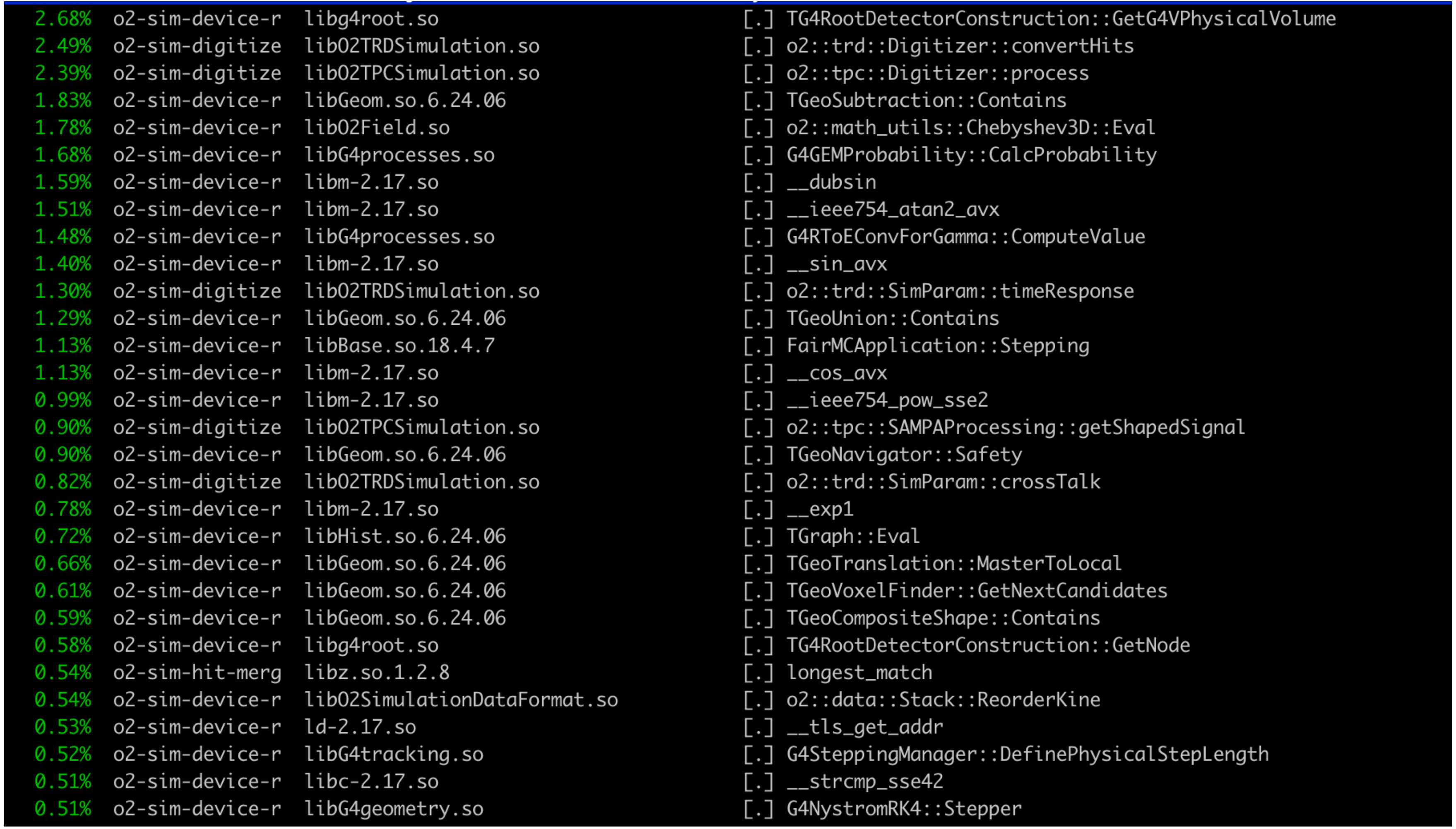
*Other names and brands may be claimed as the property of others.

Intel Confidential – Internal Use Only – Embargoed until 11 p.m. PT on Monday, May, 27.

What are we actually running?

Perf report relies on the availability of the actual code outside the container (cvmfs in our case)

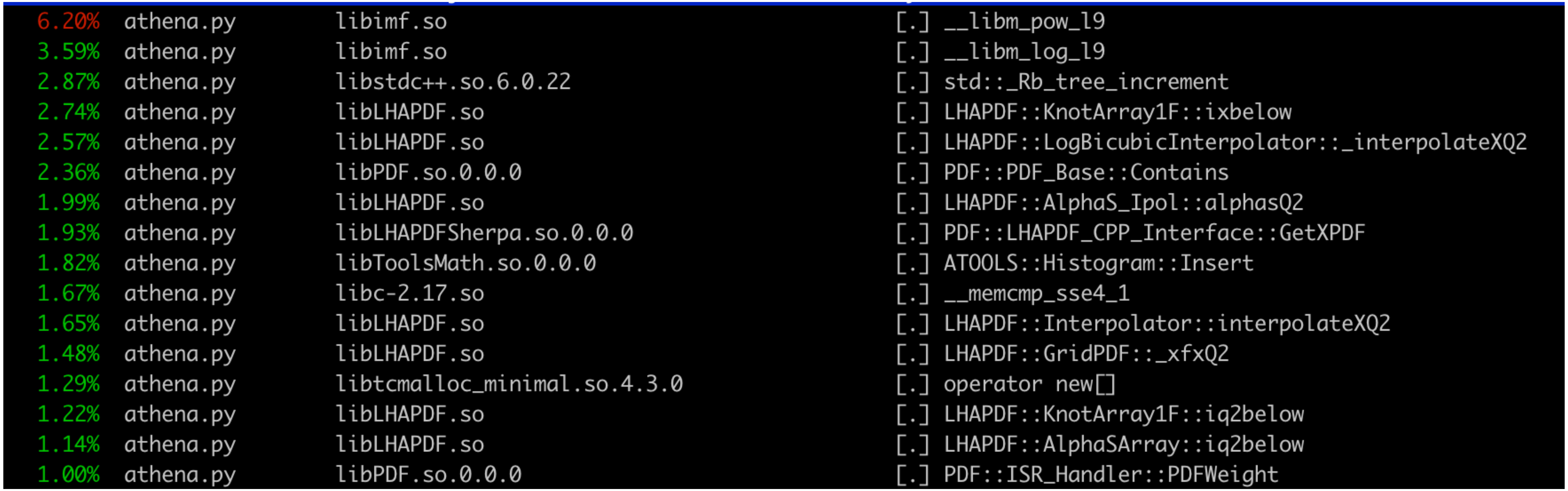
ALICE gen-sim-reco



```
2.68%  o2-sim-device-r  libg4root.so                   [.] TG4RootDetectorConstruction::GetG4VPhysicalVolume
2.49%  o2-sim-digitize  libO2TRDSimulation.so          [.] o2::trd::Digitizer::convertHits
2.39%  o2-sim-digitize  libO2TPCSimulation.so          [.] o2::tpc::Digitizer::process
1.83%  o2-sim-device-r  libGeom.so.6.24.06             [.] TGeoSubtraction::Contains
1.78%  o2-sim-device-r  libO2Field.so                  [.] o2::math_utils::Chebyshev3D::Eval
1.68%  o2-sim-device-r  libG4processes.so              [.] G4GEMProbability::CalcProbability
1.59%  o2-sim-device-r  libm-2.17.so                   [.] __dubsin
1.51%  o2-sim-device-r  libm-2.17.so                   [.] __ieee754_atan2_avx
1.48%  o2-sim-device-r  libG4processes.so              [.] G4RToEConvForGamma::ComputeValue
1.40%  o2-sim-device-r  libm-2.17.so                   [.] __sin_avx
1.30%  o2-sim-digitize  libO2TRDSimulation.so          [.] o2::trd::SimParam::timeResponse
1.29%  o2-sim-device-r  libGeom.so.6.24.06             [.] TGeoUnion::Contains
1.13%  o2-sim-device-r  libBase.so.18.4.7              [.] FairMCApplication::Stepping
1.13%  o2-sim-device-r  libm-2.17.so                   [.] __cos_avx
0.99%  o2-sim-device-r  libm-2.17.so                   [.] __ieee754_pow_sse2
0.90%  o2-sim-digitize  libO2TPCSimulation.so          [.] o2::tpc::SAMPAProcessing::getShapedSignal
0.90%  o2-sim-device-r  libGeom.so.6.24.06             [.] TGeoNavigator::Safety
0.82%  o2-sim-digitize  libO2TRDSimulation.so          [.] o2::trd::SimParam::crossTalk
0.78%  o2-sim-device-r  libm-2.17.so                   [.] __exp1
0.72%  o2-sim-device-r  libHist.so.6.24.06             [.] TGraph::Eval
0.66%  o2-sim-device-r  libGeom.so.6.24.06             [.] TGeoTranslation::MasterToLocal
0.61%  o2-sim-device-r  libGeom.so.6.24.06             [.] TGeoVoxelFinder::GetNextCandidates
0.59%  o2-sim-device-r  libGeom.so.6.24.06             [.] TGeoCompositeShape::Contains
0.58%  o2-sim-device-r  libg4root.so                   [.] TG4RootDetectorConstruction::GetNode
0.54%  o2-sim-hit-merg  libz.so.1.2.8                  [.] longest_match
0.54%  o2-sim-device-r  libO2SimulationDataFormat.so   [.] o2::data::Stack::ReorderKine
0.53%  o2-sim-device-r  ld-2.17.so                     [.] __tls_get_addr
0.52%  o2-sim-device-r  libG4tracking.so               [.] G4SteppingManager::DefinePhysicalStepLength
0.51%  o2-sim-device-r  libc-2.17.so                   [.] __strcmp_sse42
0.51%  o2-sim-device-r  libG4geometry.so               [.] G4NystromRK4::Stepper
```

- G4 Navigation
- Digitization
- libm

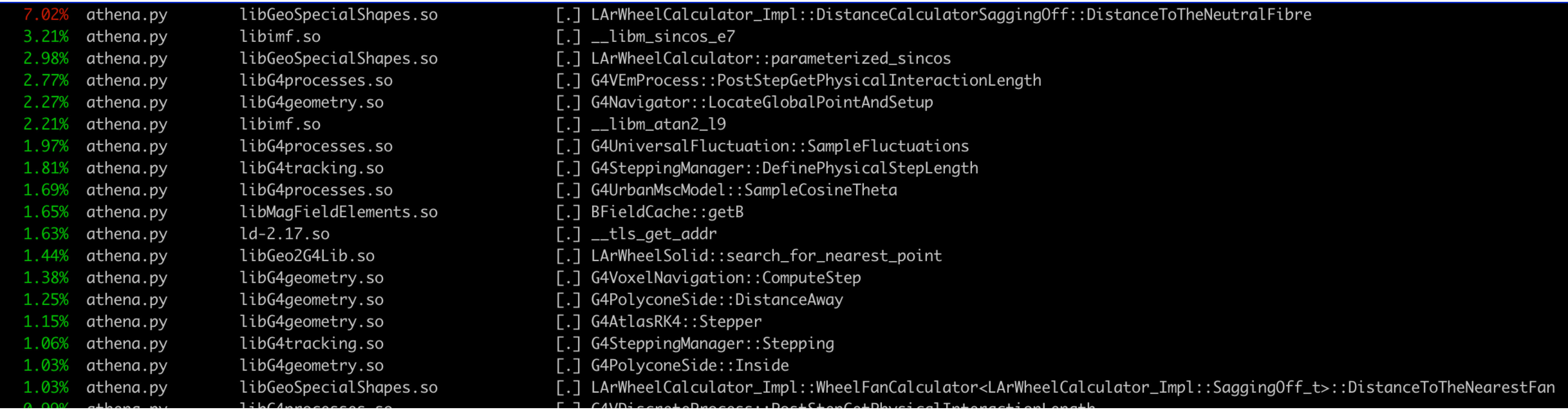
Atlas Gen (Sherpa)



```
  6.20%  athena.py       libimf.so                           [.] __libm_pow_l9
  3.59%  athena.py       libimf.so                           [.] __libm_log_l9
  2.87%  athena.py       libstdc++.so.6.0.22                 [.] std::_Rb_tree_increment
  2.74%  athena.py       libLHAPDF.so                        [.] LHAPDF::KnotArray1F::ixbelow
  2.57%  athena.py       libLHAPDF.so                        [.] LHAPDF::LogBicubicInterpolator::_interpolateXQ2
  2.36%  athena.py       libPDF.so.0.0.0                     [.] PDF::PDF_Base::Contains
  1.99%  athena.py       libLHAPDF.so                        [.] LHAPDF::AlphaS_Ipol::alphasQ2
  1.93%  athena.py       libLHAPDFSherpa.so.0.0.0            [.] PDF::LHAPDF_CPP_Interface::GetXPDF
  1.82%  athena.py       libToolsMath.so.0.0.0               [.] ATOOLS::Histogram::Insert
  1.67%  athena.py       libc-2.17.so                        [.] __memcmp_sse4_1
  1.65%  athena.py       libLHAPDF.so                        [.] LHAPDF::Interpolator::interpolateXQ2
  1.48%  athena.py       libLHAPDF.so                        [.] LHAPDF::GridPDF::_xfxQ2
  1.29%  athena.py       libtcmalloc_minimal.so.4.3.0        [.] operator new[]
  1.22%  athena.py       libLHAPDF.so                        [.] LHAPDF::KnotArray1F::iq2below
  1.14%  athena.py       libLHAPDF.so                        [.] LHAPDF::AlphaSArray::iq2below
  1.00%  athena.py       libPDF.so.0.0.0                     [.] PDF::ISR_Handler::PDFWeight
```

- INTEL mathlib
- Populating a std::map
- memcpy, tcmalloc new

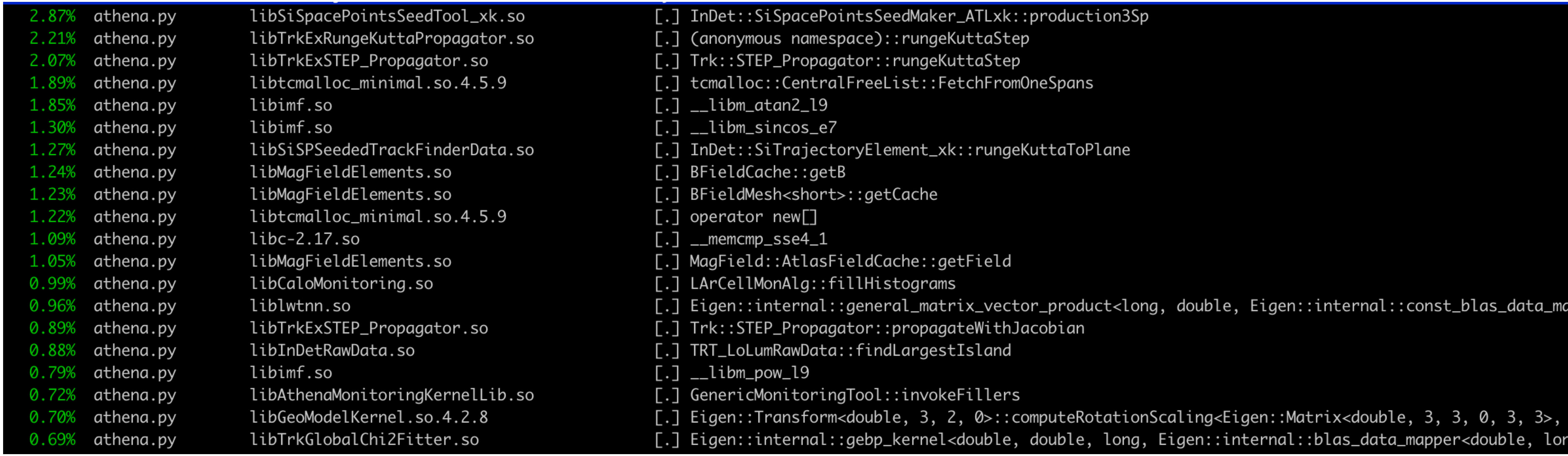
Atlas Sim



```
7.02%  athena.py         libGeoSpecialShapes.so          [.] LArWheelCalculator_Impl::DistanceCalculatorSaggingOff::DistanceToTheNeutralFibre
3.21%  athena.py         libimf.so                       [.] __libm_sincos_e7
2.98%  athena.py         libGeoSpecialShapes.so          [.] LArWheelCalculator::parameterized_sincos
2.77%  athena.py         libG4processes.so               [.] G4VEmProcess::PostStepGetPhysicalInteractionLength
2.27%  athena.py         libG4geometry.so                [.] G4Navigator::LocateGlobalPointAndSetup
2.21%  athena.py         libimf.so                       [.] __libm_atan2_l9
1.97%  athena.py         libG4processes.so               [.] G4UniversalFluctuation::SampleFluctuations
1.81%  athena.py         libG4tracking.so                [.] G4SteppingManager::DefinePhysicalStepLength
1.69%  athena.py         libG4processes.so               [.] G4UrbanMscModel::SampleCosineTheta
1.65%  athena.py         libMagFieldElements.so          [.] BFieldCache::getB
1.63%  athena.py         ld-2.17.so                      [.] __tls_get_addr
1.44%  athena.py         libGeo2G4Lib.so                 [.] LArWheelSolid::search_for_nearest_point
1.38%  athena.py         libG4geometry.so                [.] G4VoxelNavigation::ComputeStep
1.25%  athena.py         libG4geometry.so                [.] G4PolyconeSide::DistanceAway
1.15%  athena.py         libG4geometry.so                [.] G4AtlasRK4::Stepper
1.06%  athena.py         libG4tracking.so                [.] G4SteppingManager::Stepping
1.03%  athena.py         libG4geometry.so                [.] G4PolyconeSide::Inside
1.03%  athena.py         libGeoSpecialShapes.so          [.] LArWheelCalculator_Impl::WheelFanCalculator<LArWheelCalculator_Impl::SaggingOff_t>::DistanceToTheNearestFan
0.99%  athena.py         libG4processes.so               [.] G4VDiscreteProcess::PostStepGetPhysicalInteractionLength
```

- Navigation in LArWheel (including custom sincos)
- INTEL libm
- TLS management

Atlas Reco



```
  2.87%  athena.py        libSiSpacePointsSeedTool_xk.so          [.] InDet::SiSpacePointsSeedMaker_ATLxk::production3Sp
  2.21%  athena.py        libTrkExRungeKuttaPropagator.so         [.] (anonymous namespace)::rungeKuttaStep
  2.07%  athena.py        libTrkExSTEP_Propagator.so              [.] Trk::STEP_Propagator::rungeKuttaStep
  1.89%  athena.py        libtcmalloc_minimal.so.4.5.9            [.] tcmalloc::CentralFreeList::FetchFromOneSpans
  1.85%  athena.py        libimf.so                               [.] __libm_atan2_l9
  1.30%  athena.py        libimf.so                               [.] __libm_sincos_e7
  1.27%  athena.py        libSiSPSeededTrackFinderData.so         [.] InDet::SiTrajectoryElement_xk::rungeKuttaToPlane
  1.24%  athena.py        libMagFieldElements.so                  [.] BFieldCache::getB
  1.23%  athena.py        libMagFieldElements.so                  [.] BFieldMesh<short>::getCache
  1.22%  athena.py        libtcmalloc_minimal.so.4.5.9            [.] operator new[]
  1.09%  athena.py        libc-2.17.so                            [.] __memcmp_sse4_1
  1.05%  athena.py        libMagFieldElements.so                  [.] MagField::AtlasFieldCache::getField
  0.99%  athena.py        libCaloMonitoring.so                    [.] LArCellMonAlg::fillHistograms
  0.96%  athena.py        liblwtnn.so                             [.] Eigen::internal::general_matrix_vector_product<long, double, Eigen::internal::const_blas_data_ma
  0.89%  athena.py        libTrkExSTEP_Propagator.so              [.] Trk::STEP_Propagator::propagateWithJacobian
  0.88%  athena.py        libInDetRawData.so                      [.] TRT_LoLumRawData::findLargestIsland
  0.79%  athena.py        libimf.so                               [.] __libm_pow_l9
  0.72%  athena.py        libAthenaMonitoringKernelLib.so         [.] GenericMonitoringTool::invokeFillers
  0.70%  athena.py        libGeoModelKernel.so.4.2.8              [.] Eigen::Transform<double, 3, 2, 0>::computeRotationScaling<Eigen::Matrix<double, 3, 3, 0, 3, 3>,
  0.69%  athena.py        libTrkGlobalChi2Fitter.so               [.] Eigen::internal::gebp_kernel<double, double, long, Eigen::internal::blas_data_mapper<double, lon
```

- INTEL libm
- tcmalloc
- Navigation in magnetic field
- Eigen

Side remark about INTEL libm

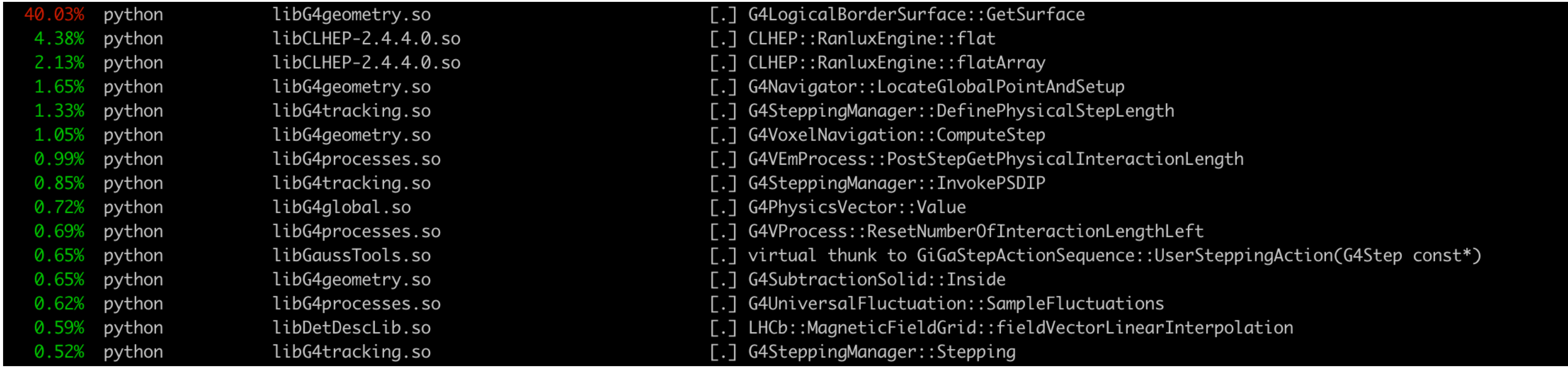
- Intel libm notoriously does not reproduce between Intel and AMD even for the very same binary code (it uses rsqrt and rcp instructions)
- Indeed some grepping in log files will show for instance that the total number of generated tracks differ

```
Intel Icelake
12:28:52 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nEvents        40
12:28:52 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nPrimaryTracks   20027
12:28:52 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nSecondaryTracks 25617
12:28:52 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO n50MeVTracks     2651431

AMD Ryzen 9 5900X
11:32:01 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nEvents        40
11:32:01 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nPrimaryTracks   20027
11:32:01 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nSecondaryTracks 24955
11:32:01 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO n50MeVTracks     2647485

Intel Haswell
15:05:59 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nEvents        40
15:05:59 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nPrimaryTracks   20027
15:05:59 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO nSecondaryTracks 25617
15:05:59 G4UA::UserActionSvc.G4UA::G4TrackCounterTool          INFO n50MeVTracks     2651431
```

https://github.com/jeff-arnold/math_routines/blob/main/rsqrt_rcp/docs/rsqrt_rcp.pdf

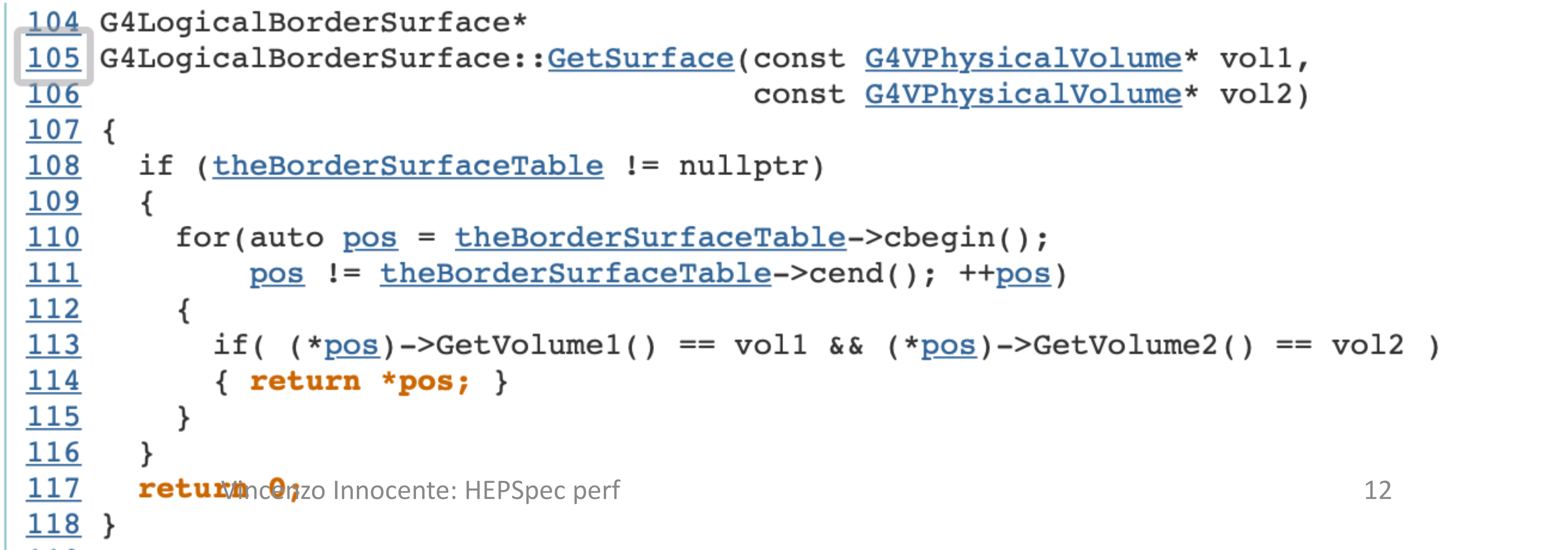
https://indico.cern.ch/event/1143946/contributions/4801434/attachments/2420216/4142495/Correctly_rounded_pow.pdf slide 8

LHCb Sim



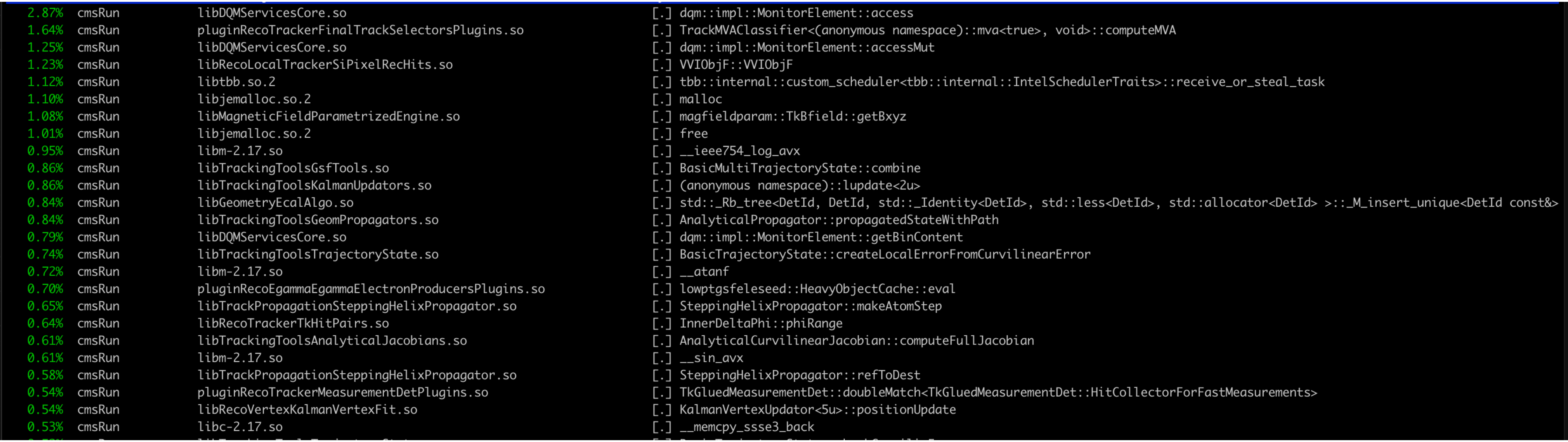
```
40.03%  python       libG4geometry.so                [.] G4LogicalBorderSurface::GetSurface
 4.38%  python       libCLHEP-2.4.4.0.so             [.] CLHEP::RanluxEngine::flat
 2.13%  python       libCLHEP-2.4.4.0.so             [.] CLHEP::RanluxEngine::flatArray
 1.65%  python       libG4geometry.so                [.] G4Navigator::LocateGlobalPointAndSetup
 1.33%  python       libG4tracking.so                [.] G4SteppingManager::DefinePhysicalStepLength
 1.05%  python       libG4geometry.so                [.] G4VoxelNavigation::ComputeStep
 0.99%  python       libG4processes.so               [.] G4VEmProcess::PostStepGetPhysicalInteractionLength
 0.85%  python       libG4tracking.so                [.] G4SteppingManager::InvokePSDIP
 0.72%  python       libG4global.so                  [.] G4PhysicsVector::Value
 0.69%  python       libG4processes.so               [.] G4VProcess::ResetNumberOfInteractionLengthLeft
 0.65%  python       libGaussTools.so                [.] virtual thunk to GiGaStepActionSequence::UserSteppingAction(G4Step const*)
 0.65%  python       libG4geometry.so                [.] G4SubtractionSolid::Inside
 0.62%  python       libG4processes.so               [.] G4UniversalFluctuation::SampleFluctuations
 0.59%  python       libDetDescLib.so                [.] LHCb::MagneticFieldGrid::fieldVectorLinearInterpolation
 0.52%  python       libG4tracking.so                [.] G4SteppingManager::Stepping
```

- Spending 40% of the time in these 4 lines of code (G4 10.6)
- code changed (from vector to map) in 10.7



```
G4LogicalBorderSurface*
G4LogicalBorderSurface::GetSurface(const G4VPhysicalVolume* vol1,
                                   const G4VPhysicalVolume* vol2)
{
  if (theBorderSurfaceTable != nullptr)
  {
    for(auto pos = theBorderSurfaceTable->cbegin();
        pos != theBorderSurfaceTable->cend(); ++pos)
    {
      if( (*pos)->GetVolume1() == vol1 && (*pos)->GetVolume2() == vol2 )
      { return *pos; }
    }
  }
  return 0;
}
```

CMS reco (DQM?)

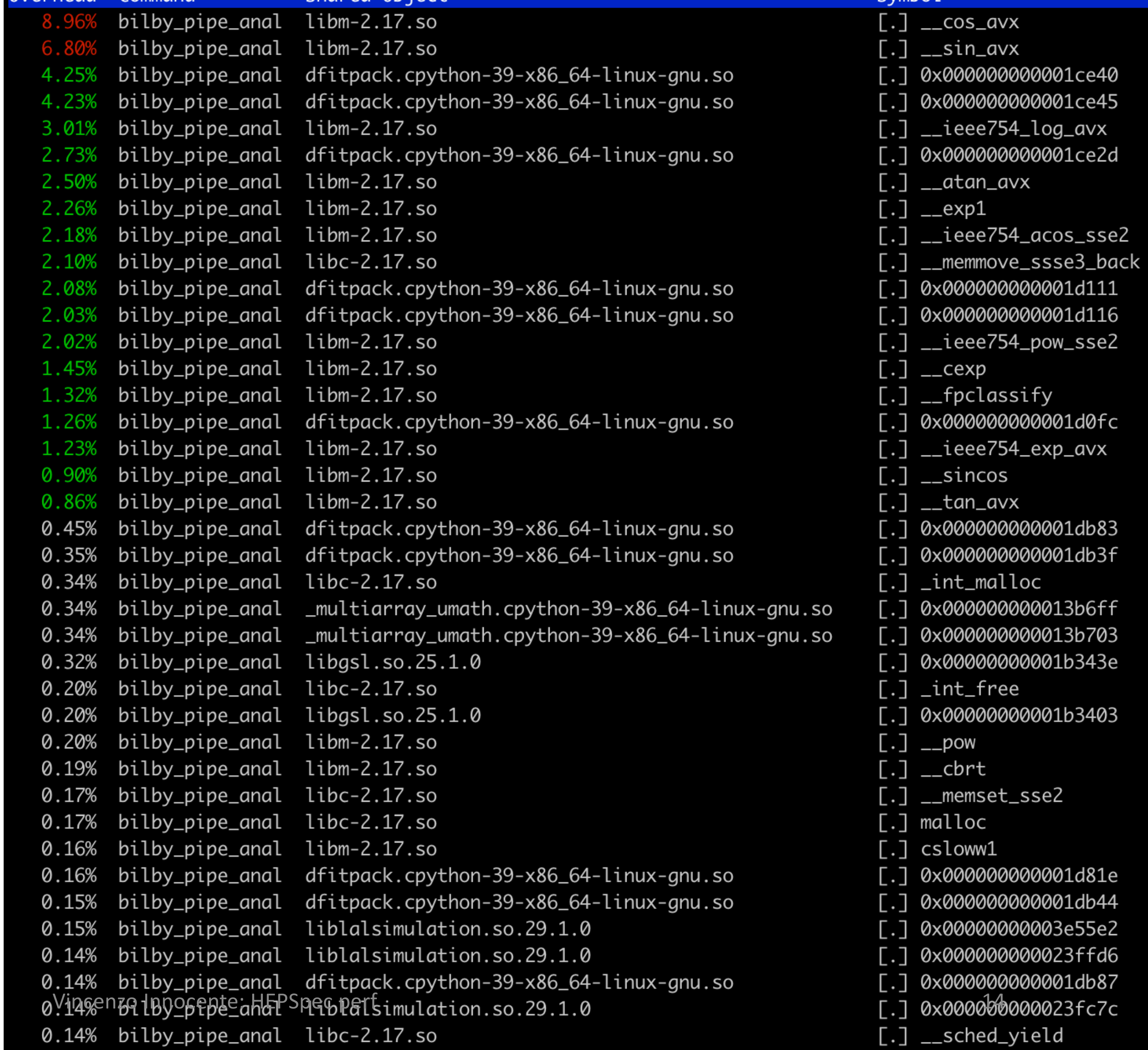


```
  2.87%  cmsRun          libDQMServicesCore.so                                [.] dqm::impl::MonitorElement::access
  1.64%  cmsRun          pluginRecoTrackerFinalTrackSelectorsPlugins.so       [.] TrackMVAClassifier<(anonymous namespace)::mva<true>, void>::computeMVA
  1.25%  cmsRun          libDQMServicesCore.so                                [.] dqm::impl::MonitorElement::accessMut
  1.23%  cmsRun          libRecoLocalTrackerSiPixelRecHits.so                 [.] VVIObjF::VVIObjF
  1.12%  cmsRun          libtbb.so.2                                          [.] tbb::internal::custom_scheduler<tbb::internal::IntelSchedulerTraits>::receive_or_steal_task
  1.10%  cmsRun          libjemalloc.so.2                                     [.] malloc
  1.08%  cmsRun          libMagneticFieldParametrizedEngine.so                [.] magfieldparam::TkBfield::getBxyz
  1.01%  cmsRun          libjemalloc.so.2                                     [.] free
  0.95%  cmsRun          libm-2.17.so                                         [.] __ieee754_log_avx
  0.86%  cmsRun          libTrackingToolsGsfTools.so                          [.] BasicMultiTrajectoryState::combine
  0.86%  cmsRun          libTrackingToolsKalmanUpdators.so                    [.] (anonymous namespace)::lupdate<2u>
  0.84%  cmsRun          libGeometryEcalAlgo.so                               [.] std::_Rb_tree<DetId, DetId, std::_Identity<DetId>, std::less<DetId>, std::allocator<DetId> >::_M_insert_unique<DetId const&>
  0.84%  cmsRun          libTrackingToolsGeomPropagators.so                   [.] AnalyticalPropagator::propagatedStateWithPath
  0.79%  cmsRun          libDQMServicesCore.so                                [.] dqm::impl::MonitorElement::getBinContent
  0.74%  cmsRun          libTrackingToolsTrajectoryState.so                   [.] BasicTrajectoryState::createLocalErrorFromCurvilinearError
  0.72%  cmsRun          libm-2.17.so                                         [.] __atanf
  0.70%  cmsRun          pluginRecoEgammaEgammaElectronProducersPlugins.so    [.] lowptgsfeleseed::HeavyObjectCache::eval
  0.65%  cmsRun          libTrackPropagationSteppingHelixPropagator.so        [.] SteppingHelixPropagator::makeAtomStep
  0.64%  cmsRun          libRecoTrackerTkHitPairs.so                          [.] InnerDeltaPhi::phiRange
  0.61%  cmsRun          libTrackingToolsAnalyticalJacobians.so               [.] AnalyticalCurvilinearJacobian::computeFullJacobian
  0.61%  cmsRun          libm-2.17.so                                         [.] __sin_avx
  0.58%  cmsRun          libTrackPropagationSteppingHelixPropagator.so        [.] SteppingHelixPropagator::refToDest
  0.54%  cmsRun          pluginRecoTrackerMeasurementDetPlugins.so            [.] TkGluedMeasurementDet::doubleMatch<TkGluedMeasurementDet::HitCollectorForFastMeasurements>
  0.54%  cmsRun          libRecoVertexKalmanVertexFit.so                      [.] KalmanVertexUpdator<5u>::positionUpdate
  0.53%  cmsRun          libc-2.17.so                                         [.] __memcpy_ssse3_back
```

- Filling histograms
- Tracking MVA
- tbb overhead
- malloc/free
- Magnetic field

igwn

- Libm (almost a hot spot)
- Other libraries not available "on metal"



```
 8.96%  bilby_pipe_anal  libm-2.17.so                                        [.] __cos_avx
 6.80%  bilby_pipe_anal  libm-2.17.so                                        [.] __sin_avx
 4.25%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001ce40
 4.23%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001ce45
 3.01%  bilby_pipe_anal  libm-2.17.so                                        [.] __ieee754_log_avx
 2.73%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001ce2d
 2.50%  bilby_pipe_anal  libm-2.17.so                                        [.] __atan_avx
 2.26%  bilby_pipe_anal  libm-2.17.so                                        [.] __exp1
 2.18%  bilby_pipe_anal  libm-2.17.so                                        [.] __ieee754_acos_sse2
 2.10%  bilby_pipe_anal  libc-2.17.so                                        [.] __memmove_ssse3_back
 2.08%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001d111
 2.03%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001d116
 2.02%  bilby_pipe_anal  libm-2.17.so                                        [.] __ieee754_pow_sse2
 1.45%  bilby_pipe_anal  libm-2.17.so                                        [.] __cexp
 1.32%  bilby_pipe_anal  libm-2.17.so                                        [.] __fpclassify
 1.26%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001d0fc
 1.23%  bilby_pipe_anal  libm-2.17.so                                        [.] __ieee754_exp_avx
 0.90%  bilby_pipe_anal  libm-2.17.so                                        [.] __sincos
 0.86%  bilby_pipe_anal  libm-2.17.so                                        [.] __tan_avx
 0.45%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001db83
 0.35%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001db3f
 0.34%  bilby_pipe_anal  libc-2.17.so                                        [.] _int_malloc
 0.34%  bilby_pipe_anal  _multiarray_umath.cpython-39-x86_64-linux-gnu.so    [.] 0x000000000013b6ff
 0.34%  bilby_pipe_anal  _multiarray_umath.cpython-39-x86_64-linux-gnu.so    [.] 0x000000000013b703
 0.32%  bilby_pipe_anal  libgsl.so.25.1.0                                    [.] 0x00000000001b343e
 0.20%  bilby_pipe_anal  libc-2.17.so                                        [.] _int_free
 0.20%  bilby_pipe_anal  libgsl.so.25.1.0                                    [.] 0x00000000001b3403
 0.20%  bilby_pipe_anal  libm-2.17.so                                        [.] __pow
 0.19%  bilby_pipe_anal  libm-2.17.so                                        [.] __cbrt
 0.17%  bilby_pipe_anal  libc-2.17.so                                        [.] __memset_sse2
 0.17%  bilby_pipe_anal  libc-2.17.so                                        [.] malloc
 0.16%  bilby_pipe_anal  libm-2.17.so                                        [.] csloww1
 0.16%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001d81e
 0.15%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001db44
 0.15%  bilby_pipe_anal  liblalsimulation.so.29.1.0                          [.] 0x00000000003e55e2
 0.14%  bilby_pipe_anal  liblalsimulation.so.29.1.0                          [.] 0x000000000023ffd6
 0.14%  bilby_pipe_anal  dfitpack.cpython-39-x86_64-linux-gnu.so             [.] 0x000000000001db87
 0.14%  bilby_pipe_anal  liblalsimulation.so.29.1.0                          [.] 0x000000000023fc7c
 0.14%  bilby_pipe_anal  libc-2.17.so                                        [.] __sched_yield
```

Summary (1)

- No notable hotspot
- Major exception
  - Vector scan in LHCb simulation (may change in future versions)
- Minor exceptions
  - libm in igwn and sherpa (may change in future OS or using alternative libm)
  - Histogram filling in CMS reco
  - “Navigation” in LArWheel for Atlas Sim

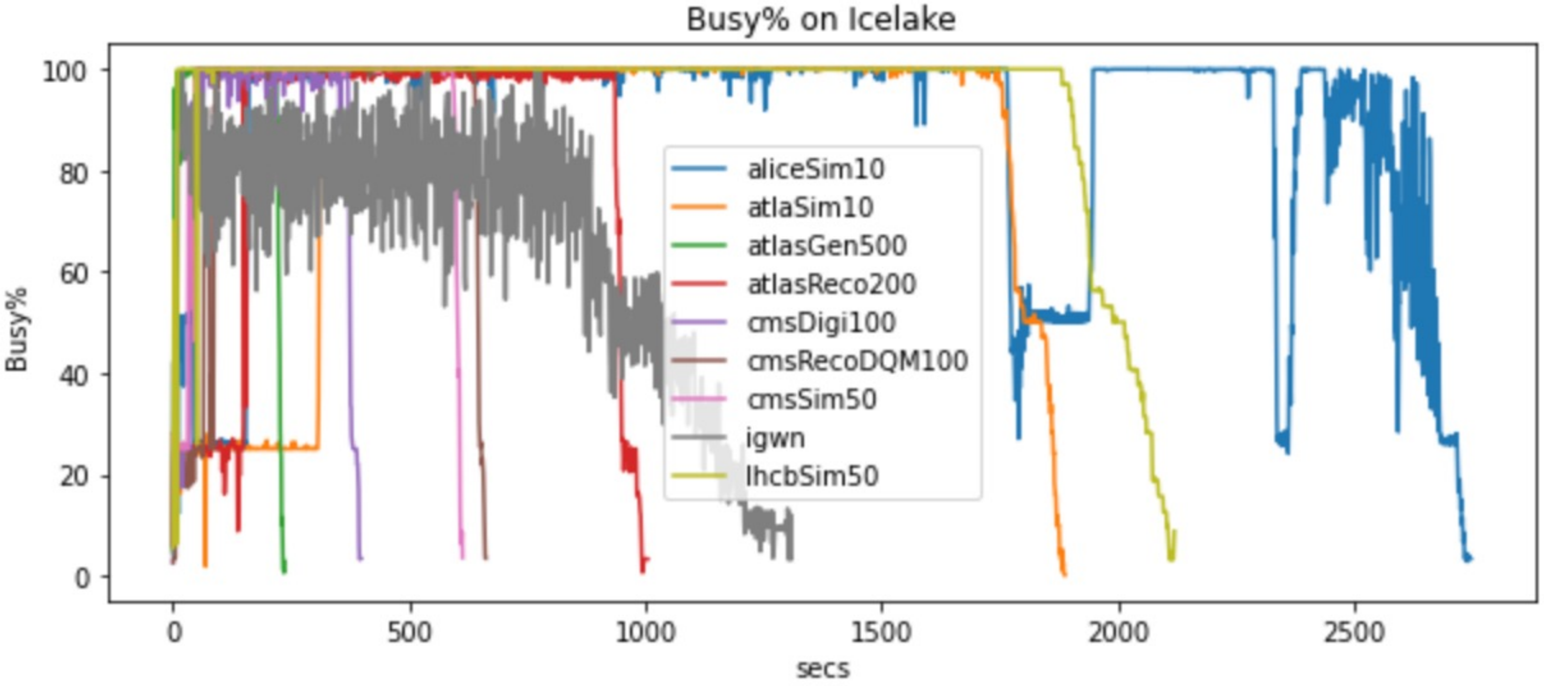
Resource Utilization

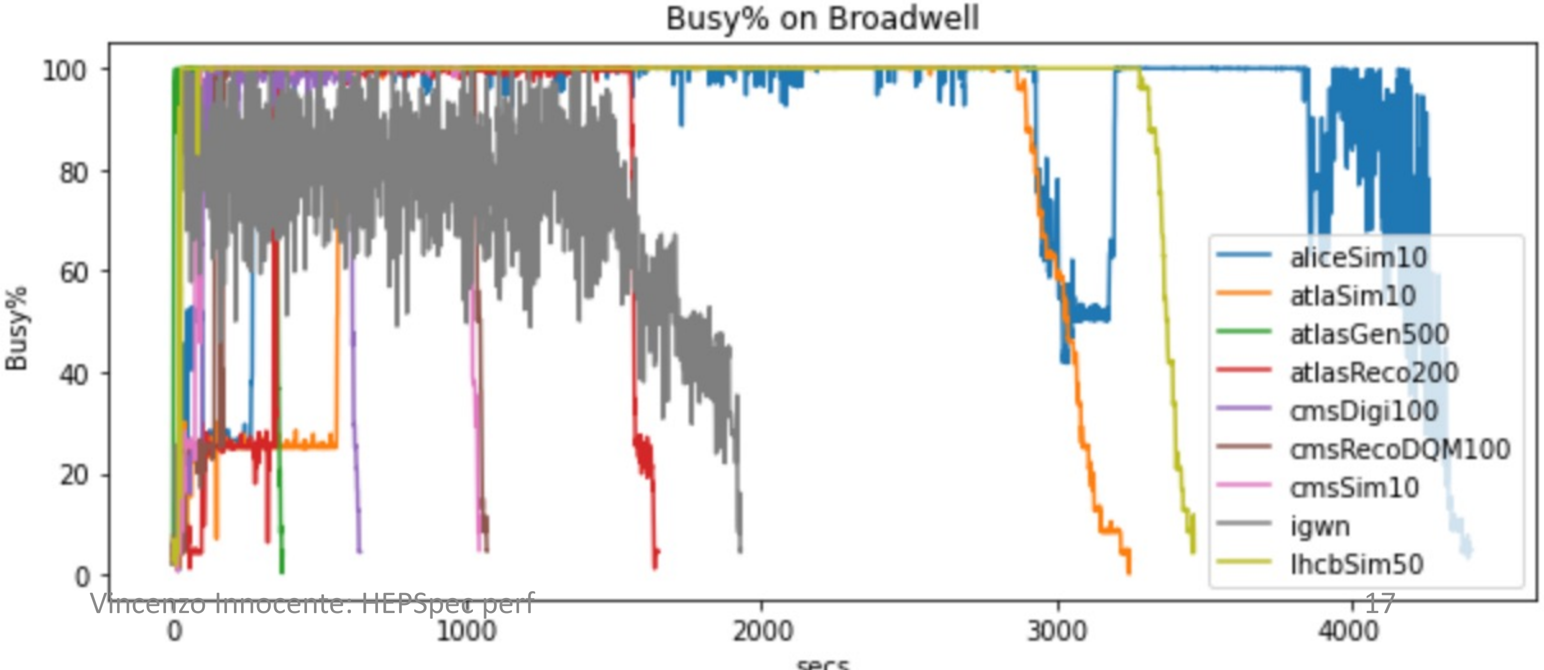
(full machine, no HT unless explicitly quoted)

turbostat

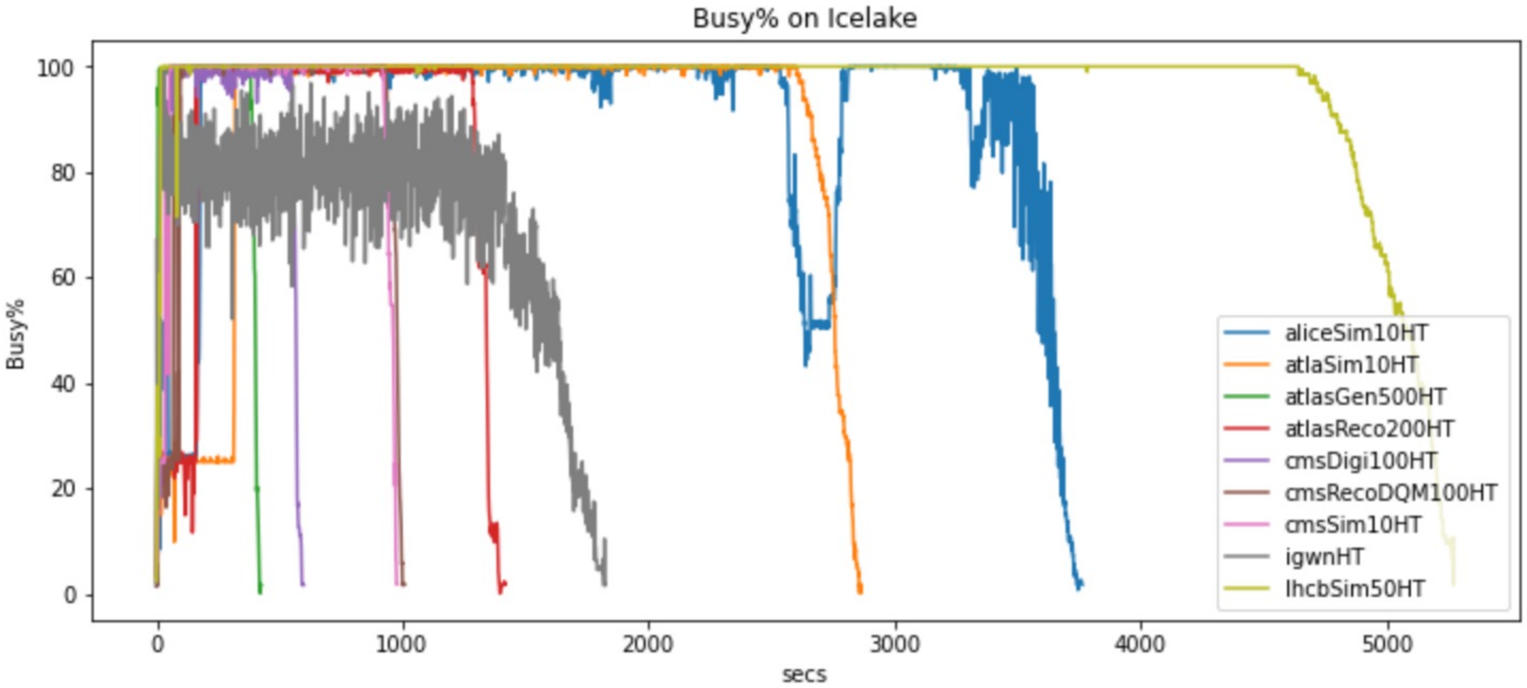
```
turbostat --interval=1 -S >& aliceSim10F.turbolog&
./doPerfStat "singularity run -B $workdir:/results oras://registry.cern.ch/hep-workloads/alice-gen-sim-reco-run3-bmk:ci-v0.6-aod -t4 -e10" >& aliceSim10F.perflog
killall -9 turbostat
```

Busy %

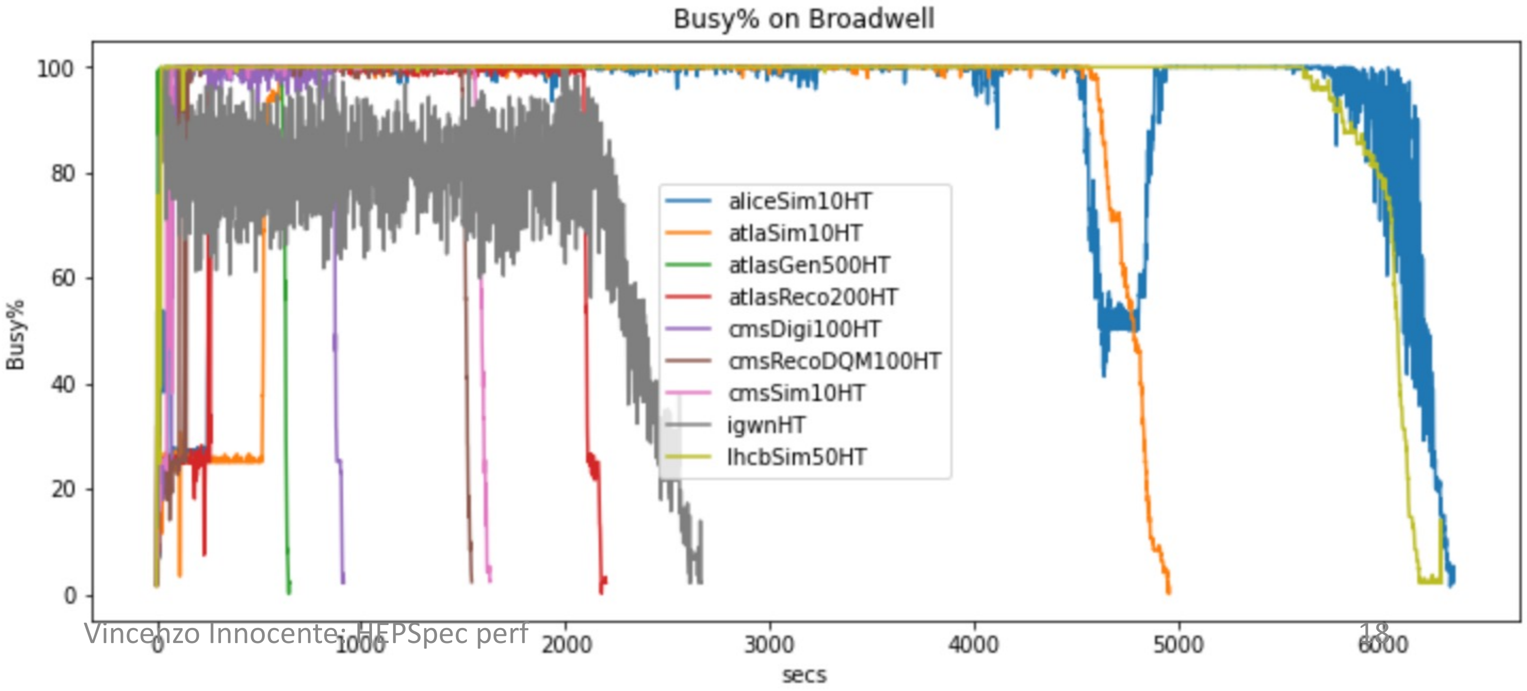

Busy% on Icelake
Busy%
100
80
60
40
20
0
0
500
1000
1500
2000
2500
secs
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim50
igwn
lhcbSim50


Busy% on Broadwell
Busy%
100
80
60
40
20
0
0
1000
2000
3000
4000
secs
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim10
igwn
lhcbSim50

Busy % (HT on)

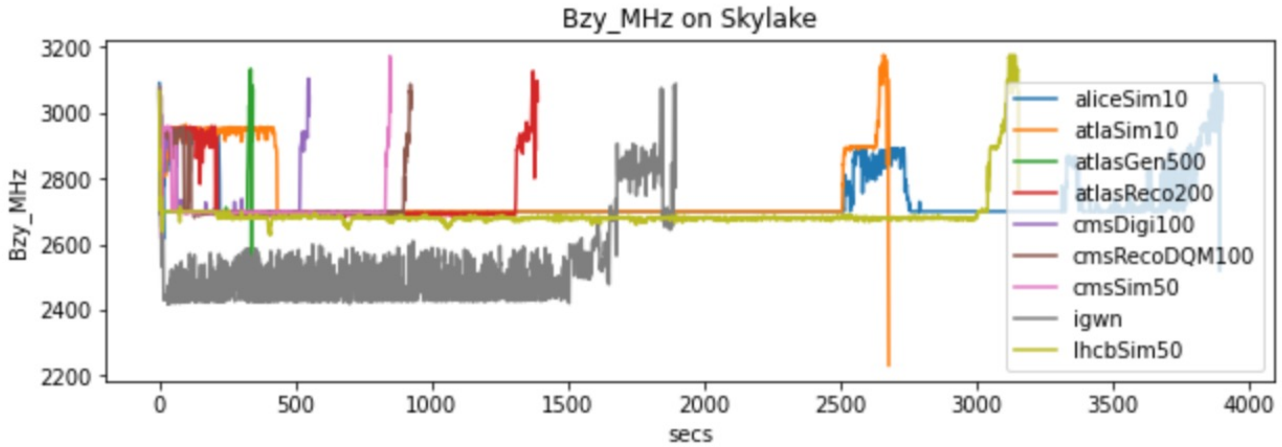


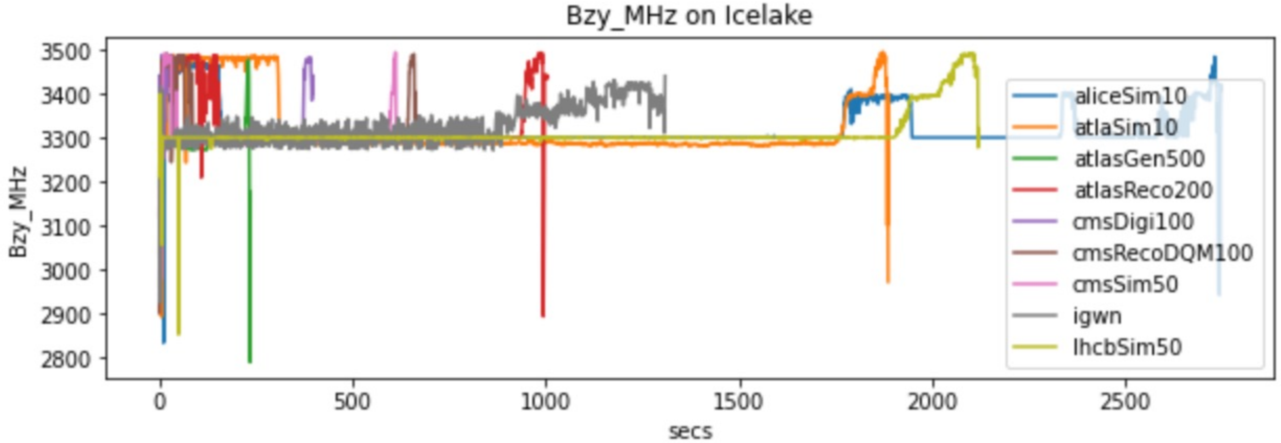
Busy% on Icelake
Busy%
100
80
60
40
20
0
0
1000
2000
3000
4000
5000
secs
aliceSim10HT
atlaSim10HT
atlasGen500HT
atlasReco200HT
cmsDigi100HT
cmsRecoDQM100HT
cmsSim10HT
igwnHT
lhcbSim50HT

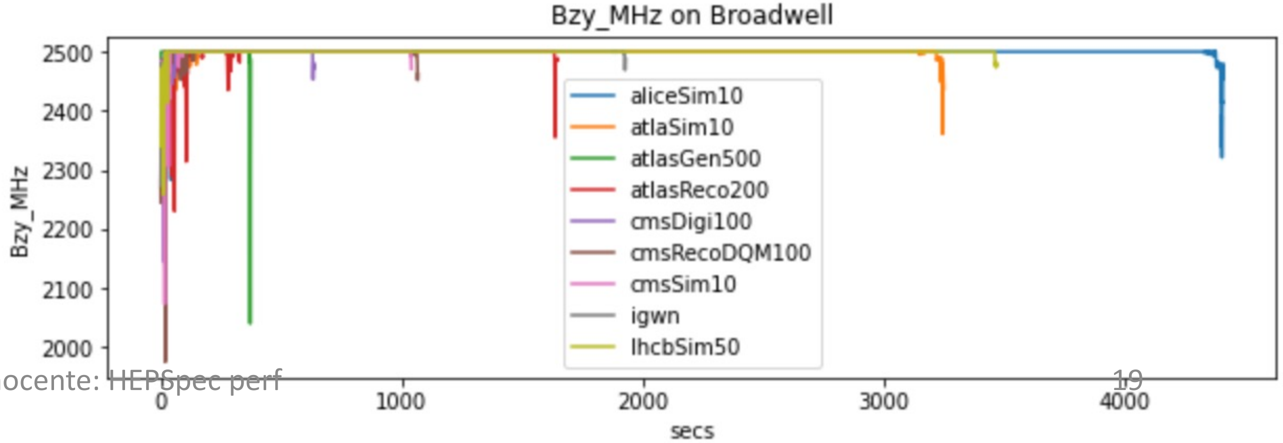


Busy% on Broadwell
Busy%
100
80
60
40
20
0
0
1000
2000
3000
4000
5000
6000
secs
aliceSim10HT
atlaSim10HT
atlasGen500HT
atlasReco200HT
cmsDigi100HT
cmsRecoDQM100HT
cmsSim10HT
igwnHT
lhcbSim50HT

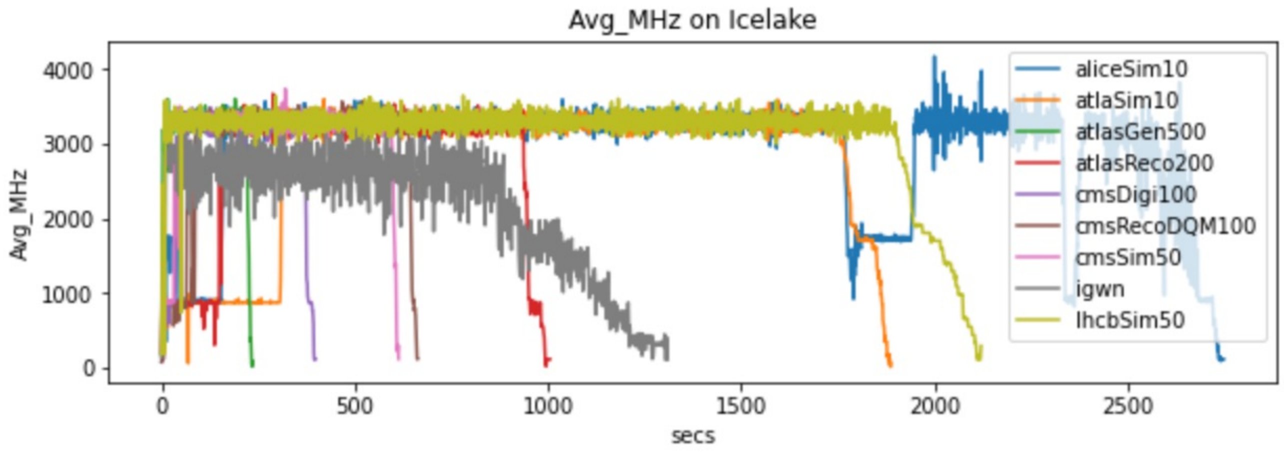
Busy Freq

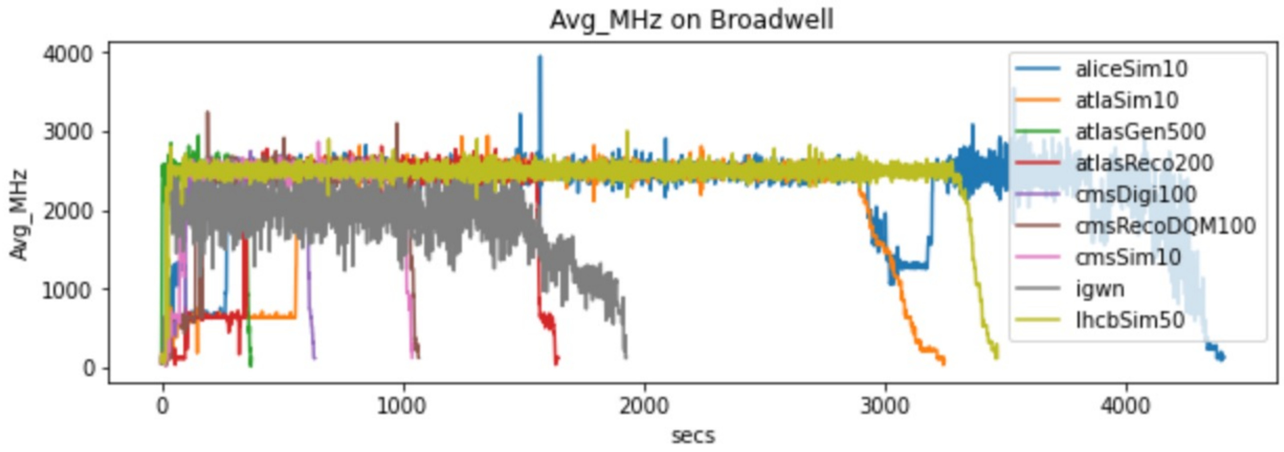

Bzy_MHz on Skylake

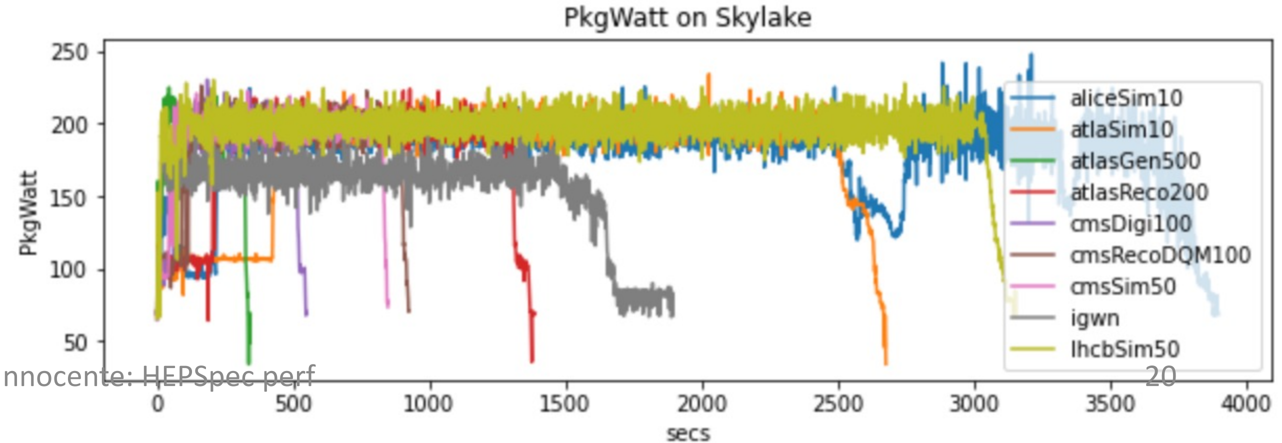

Bzy_MHz on Icelake


Bzy_MHz on Broadwell

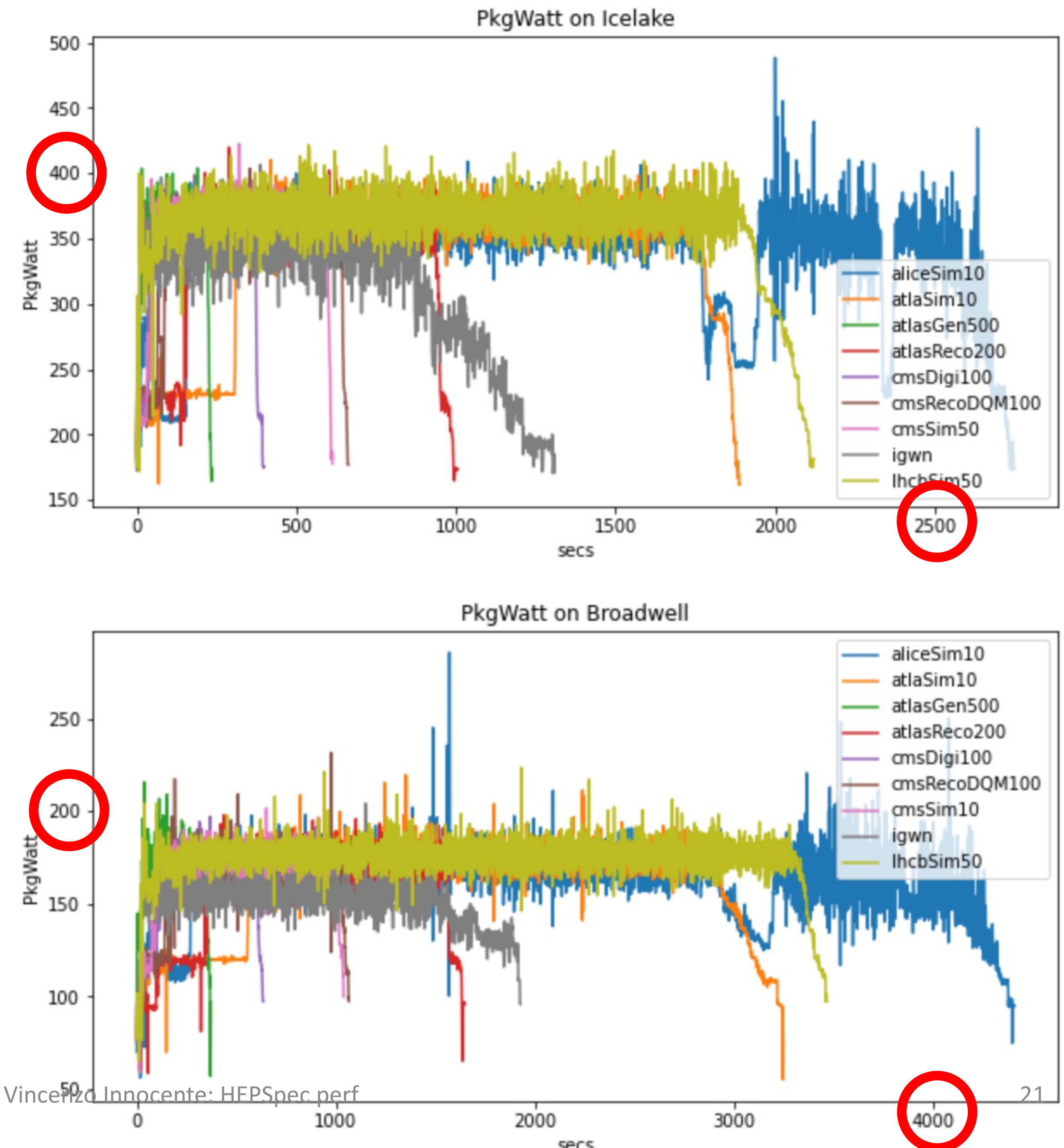
Ave Freq


Avg_MHz on Icelake

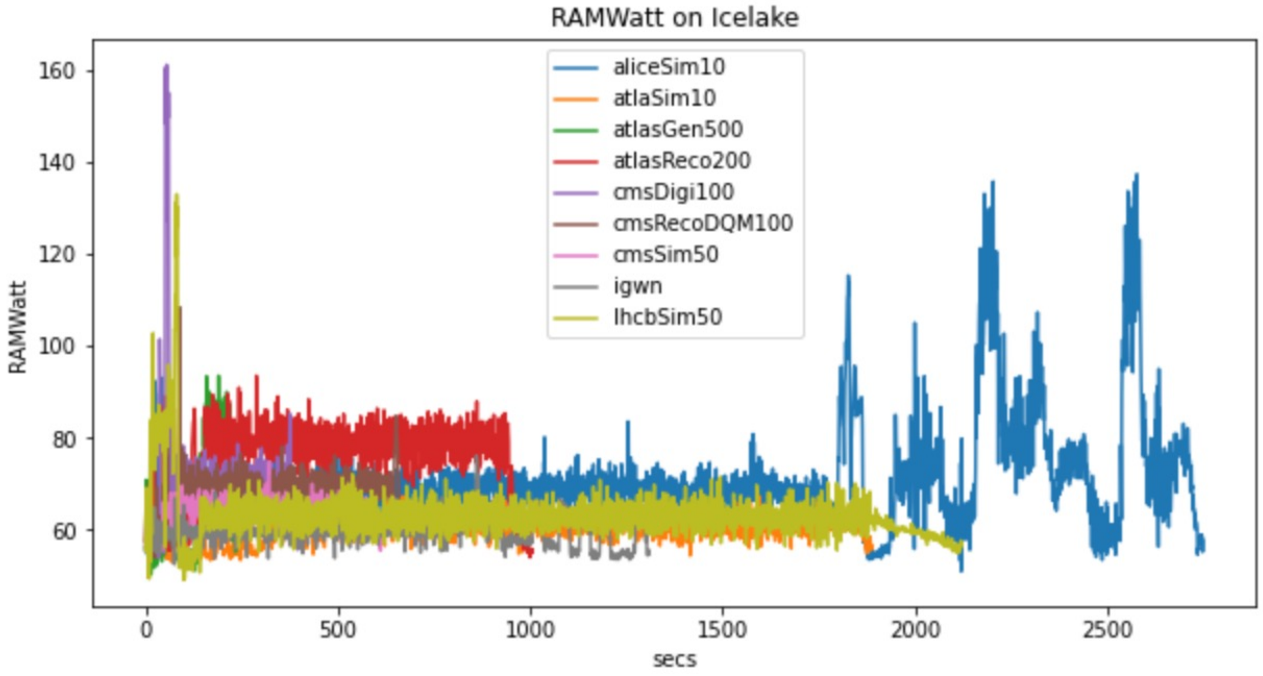

Avg_MHz on Broadwell

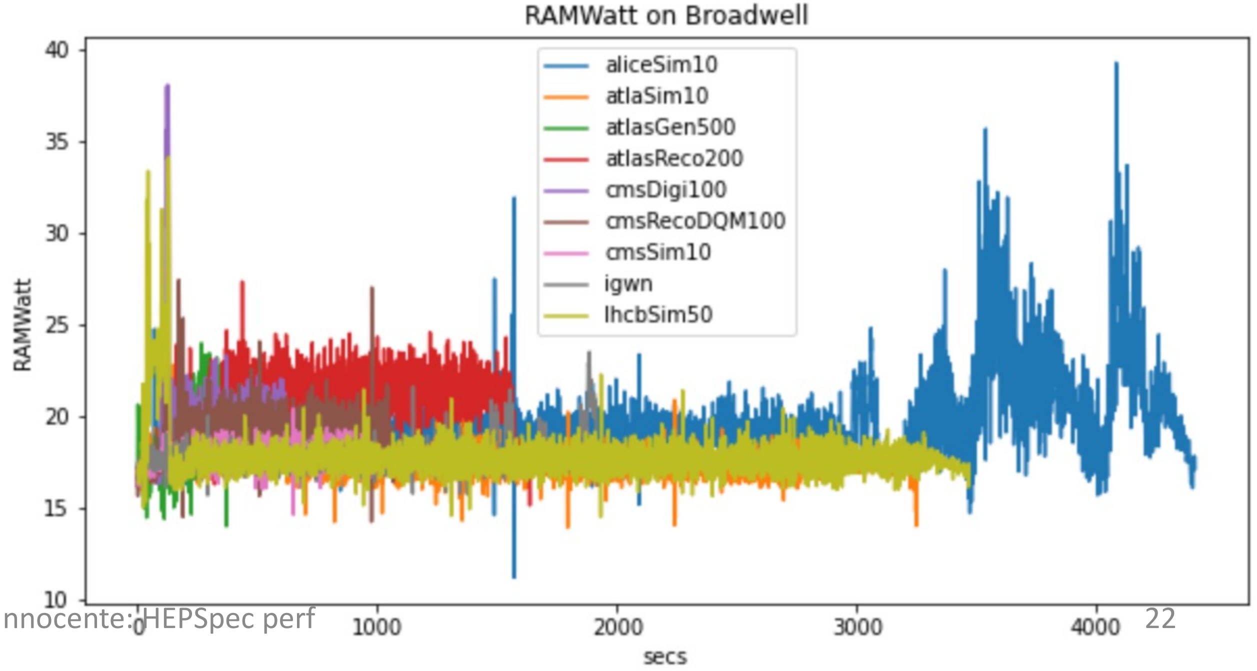

PkgWatt on Skylake

Core Power

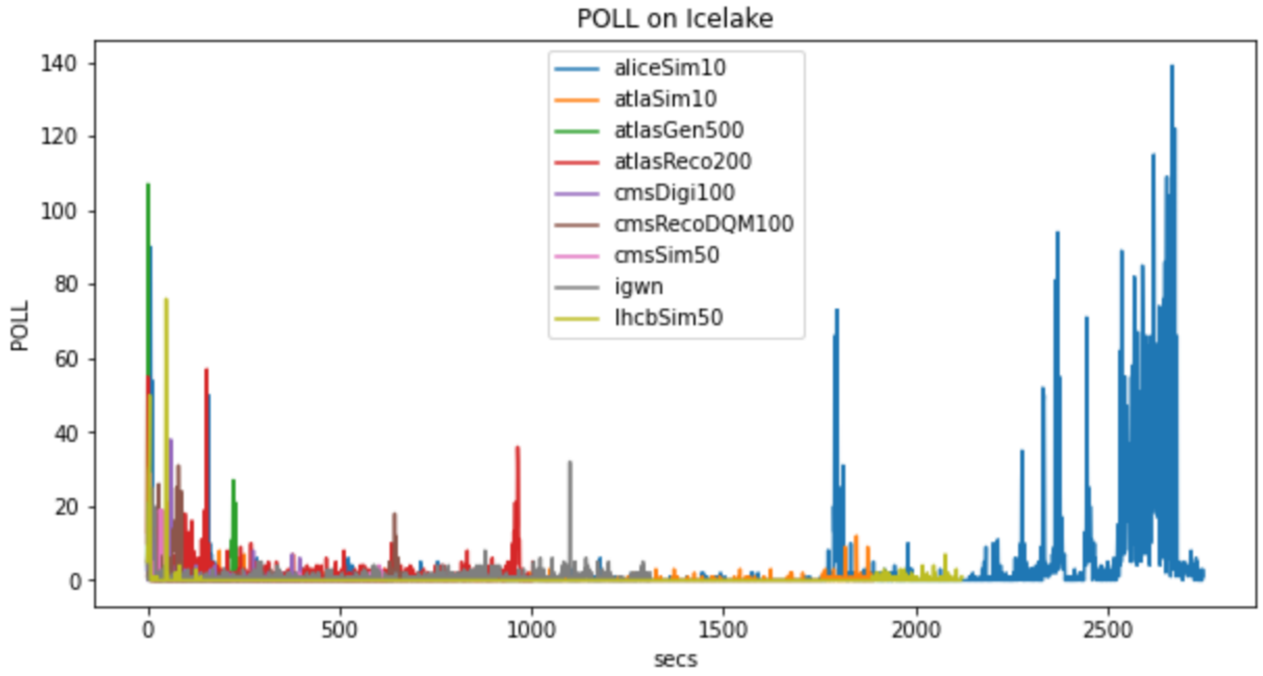

PkgWatt on Icelake
PkgWatt
500
450
400
350
300
250
200
150
0
500
1000
1500
2000
2500
secs
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim50
igwn
lhcbSim50
PkgWatt on Broadwell
PkgWatt
250
200
150
100
50
0
1000
2000
3000
4000
secs
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim10
igwn
lhcbSim50

Memory Power

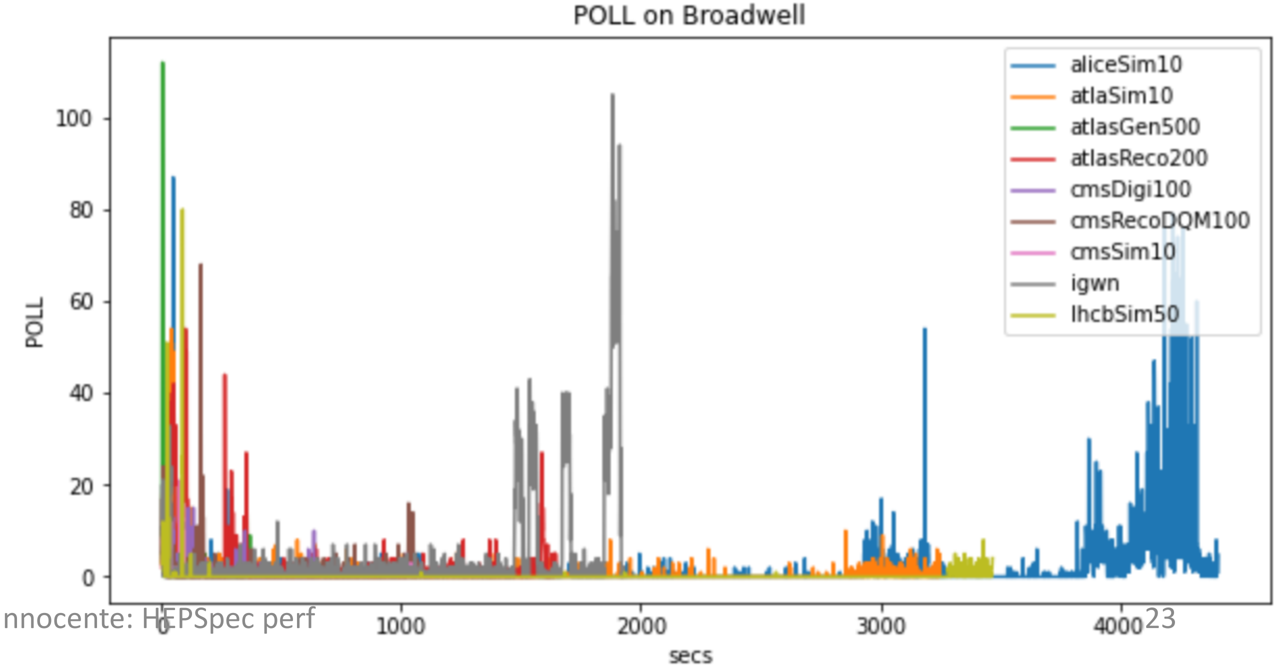




Poll (I/O)

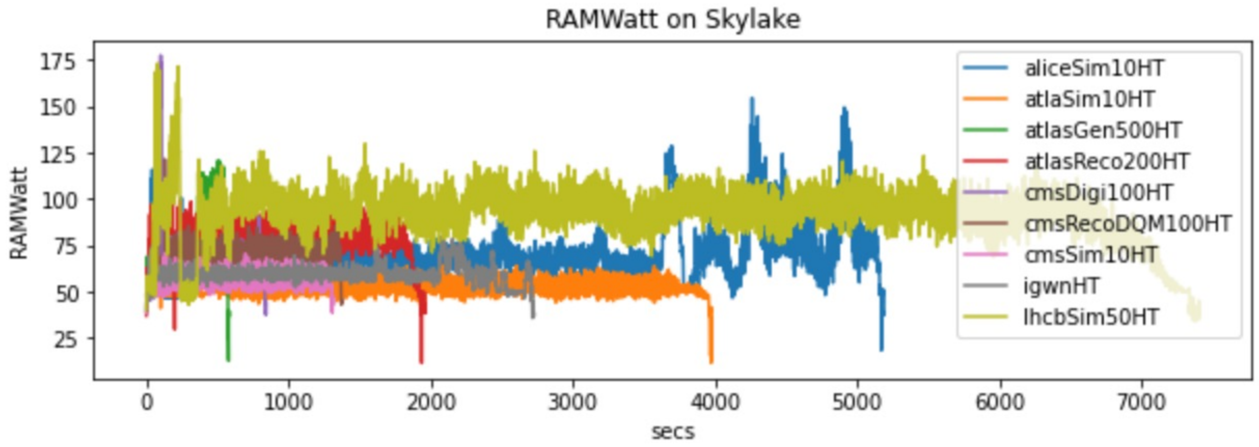


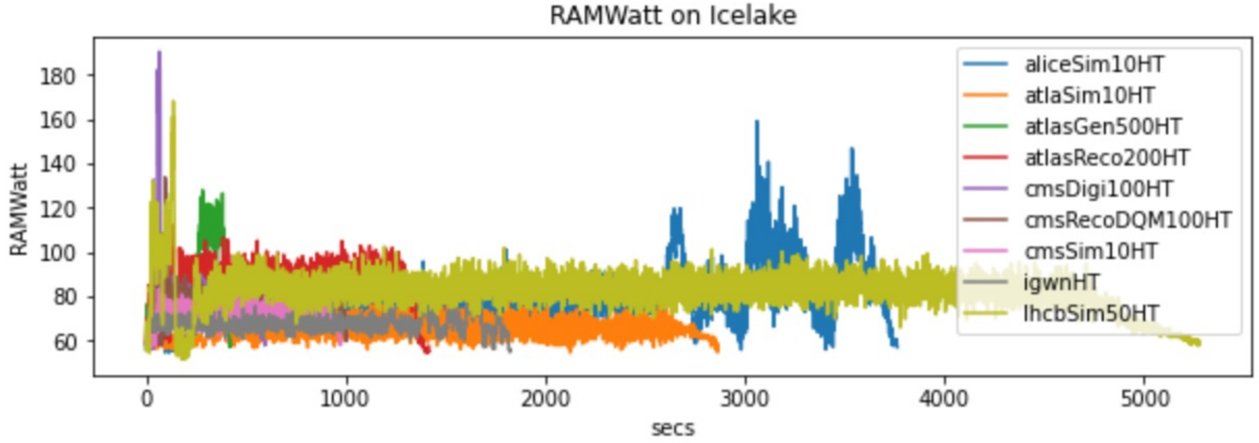
POLL on Icelake
POLL
secs
0
20
40
60
80
100
120
140
0
500
1000
1500
2000
2500
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim50
igwn
lhcbSim50

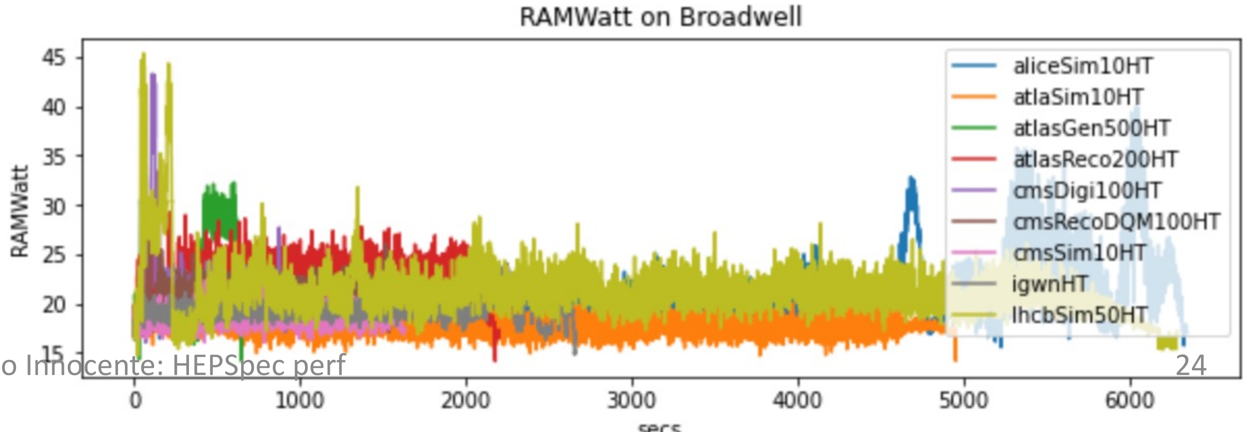


POLL on Broadwell
POLL
secs
0
20
40
60
80
100
0
1000
2000
3000
4000
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim10
igwn
lhcbSim50

Memory Power (HT on)







Summary (2)

- All wf but igwn keep cores 100% busy during main processing
- Igwn uses 80% of the cores and shows a long ending tail
  - Can be mitigated by over-committing
- Alice wf has long low efficiency stretches and a tail of low cpu efficiency and high memory and I/O usage.
- ATLAS reco shows large memory access during processing
- ATLAS sim and reco shows “long” single thread initialization
  - Long production job will not be affected by that
- LHCb memory usage is very high if HT is on (affecting timing)
- Igwn seems to suffer from freq-throttling on Skylake

Detailed Perf Statistics

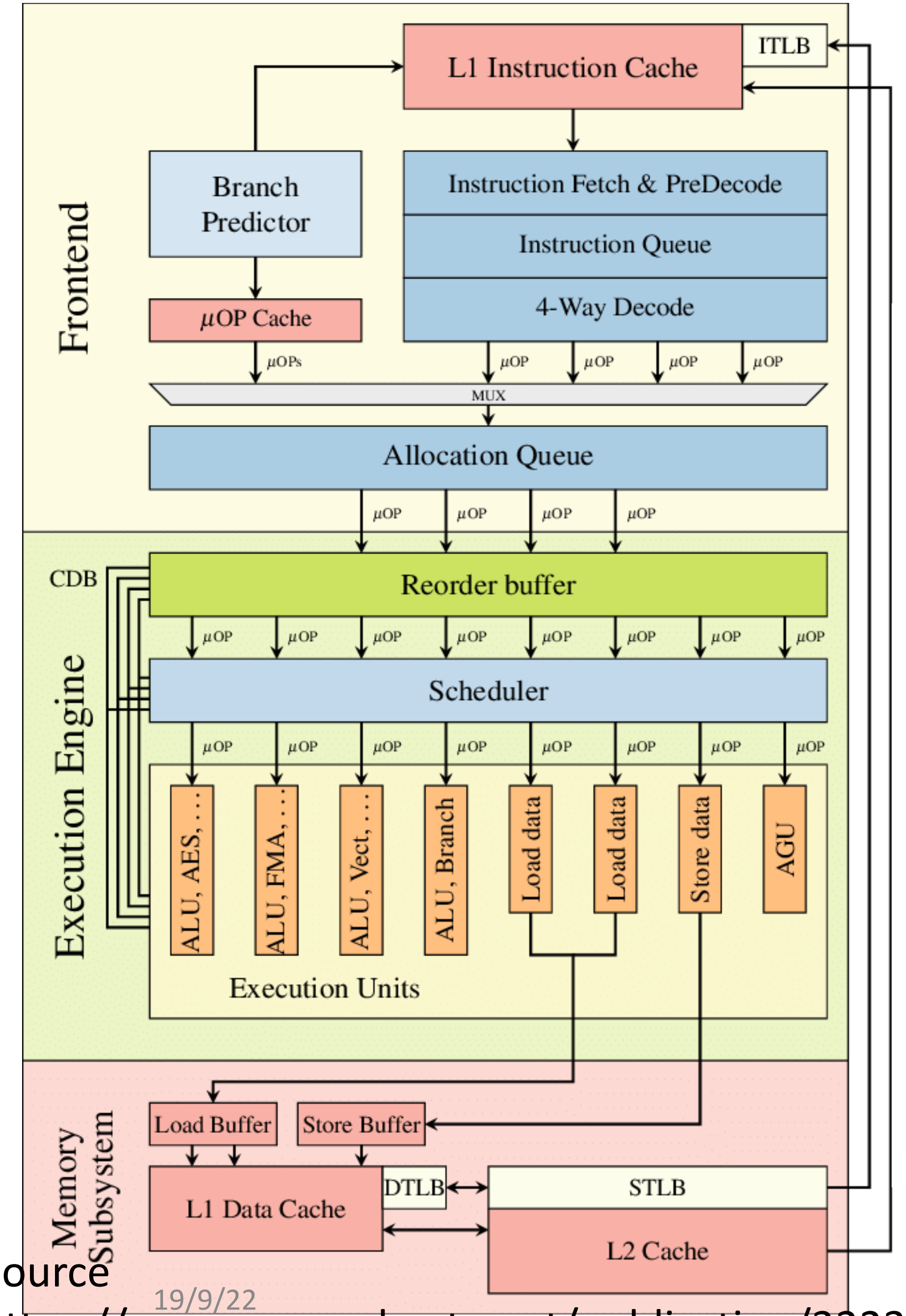
Full Machine

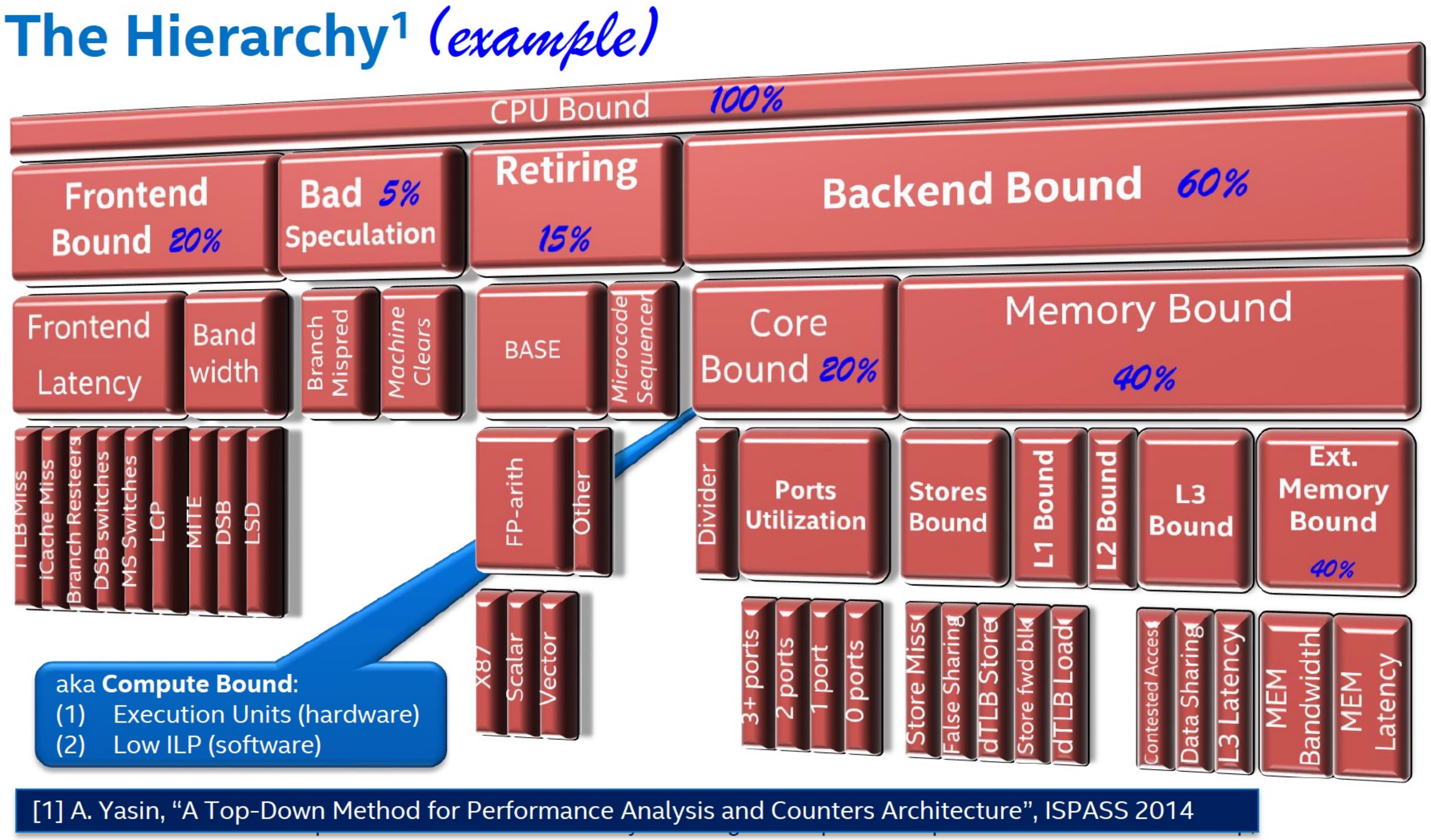
No HT unless specified

Guilherme Amadio @SFT meeting https://indico.cern.ch/event/980497/

Vincenzo Innocente @ESC18 https://agenda.infn.it/event/16941/contributions/34860/attachments/24525/27968/Architecture_ESC18.pdf

Reminder: Processor Architecture





The Hierarchy[1] *(example)*

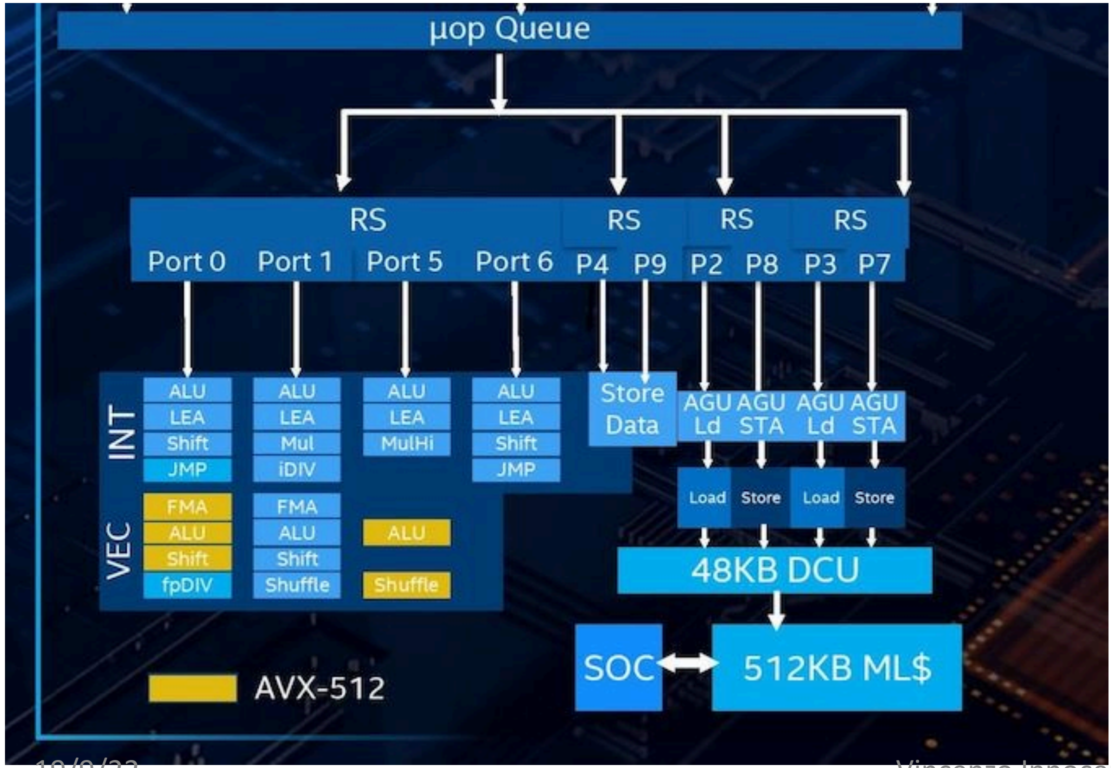
aka **Compute Bound**:
(1) Execution Units (hardware)
(2) Low ILP (software)

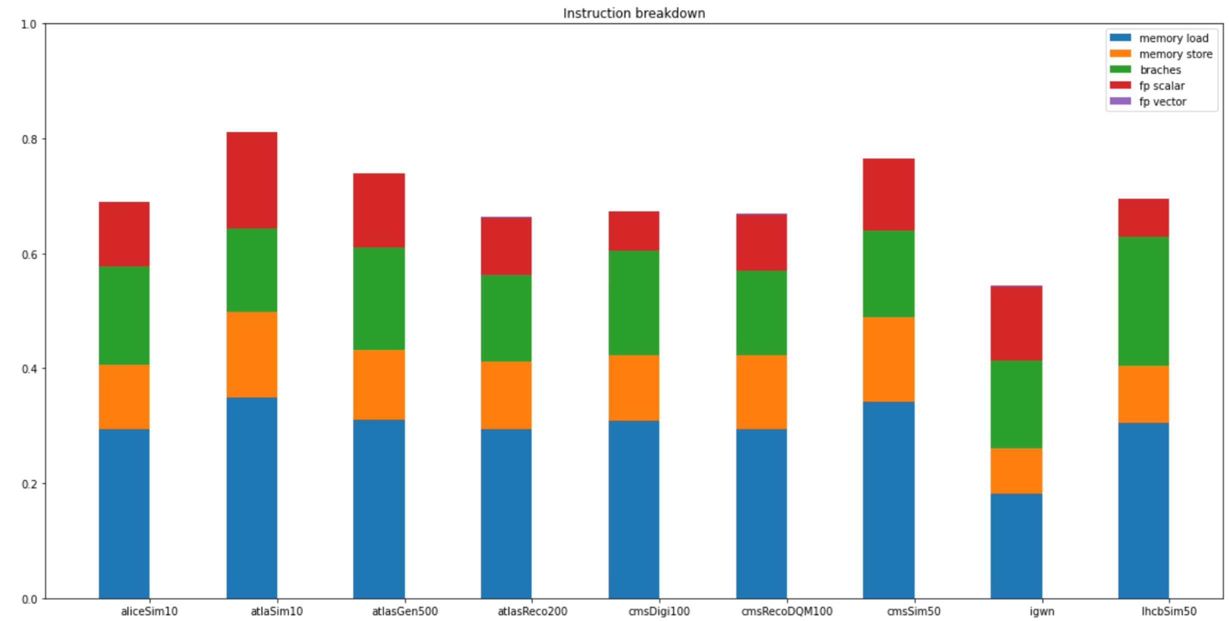
[1] A. Yasin, "A Top-Down Method for Performance Analysis and Counters Architecture", ISPASS 2014

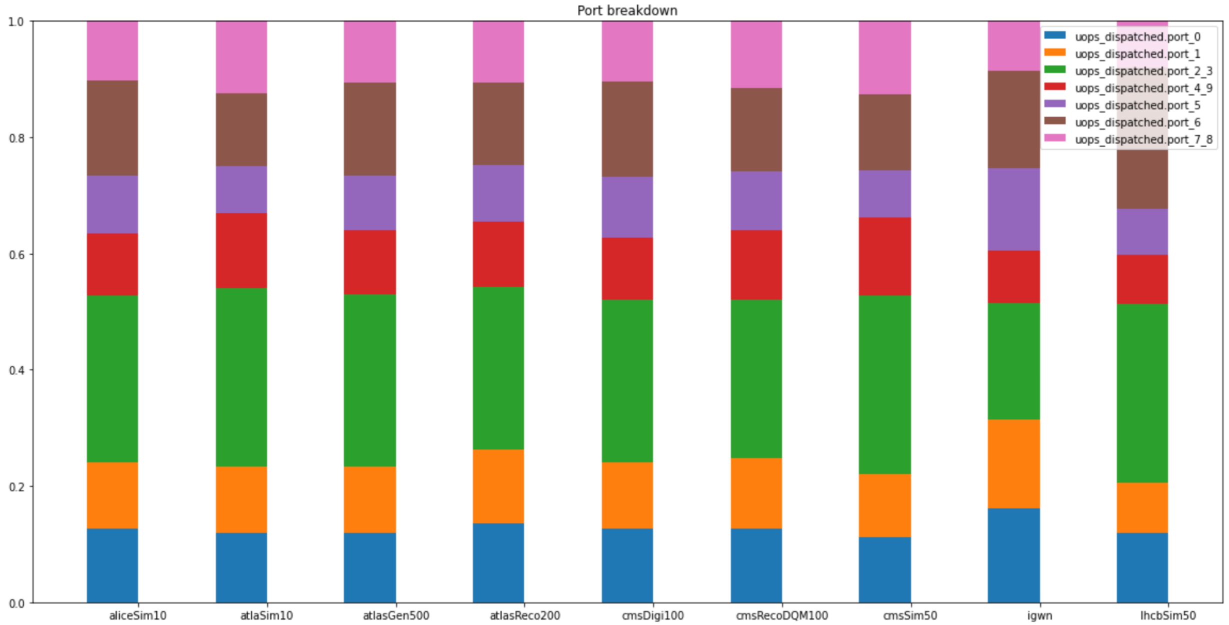
17

Source https://www.researchgate.net/publication/322305793_Meltdown

Instruction Breakdown (on Icelake)

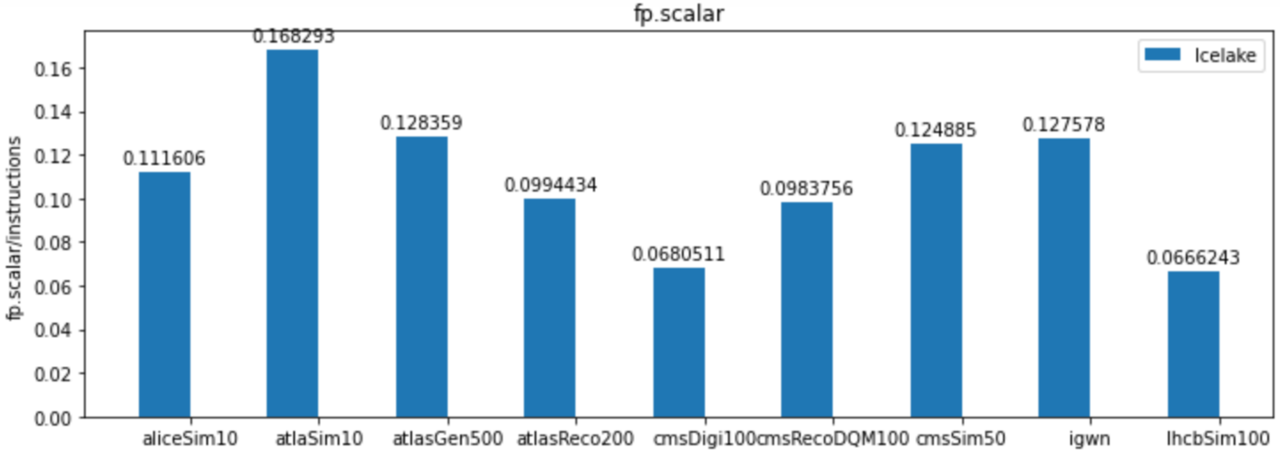


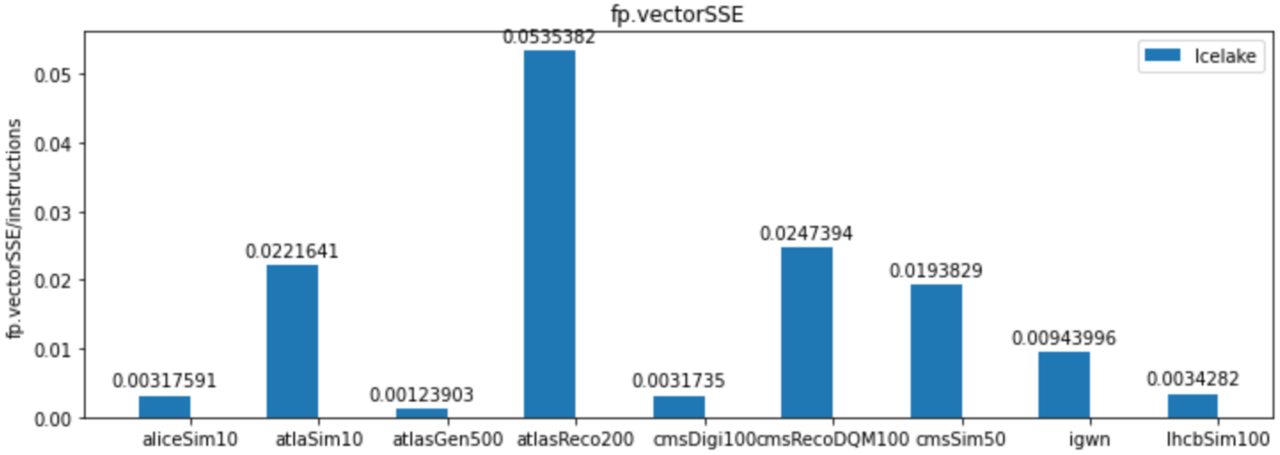


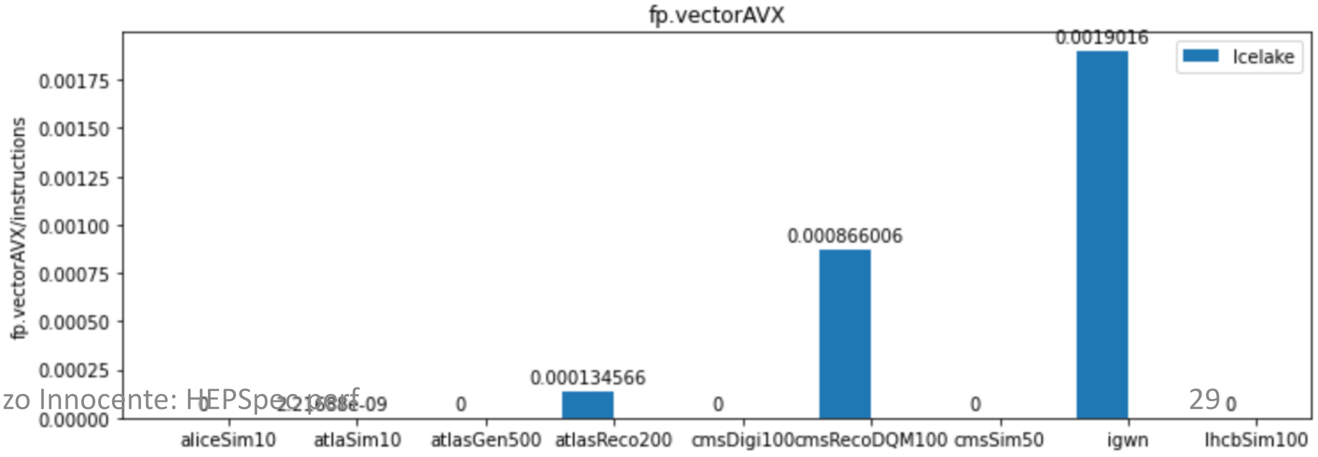


Floating-point

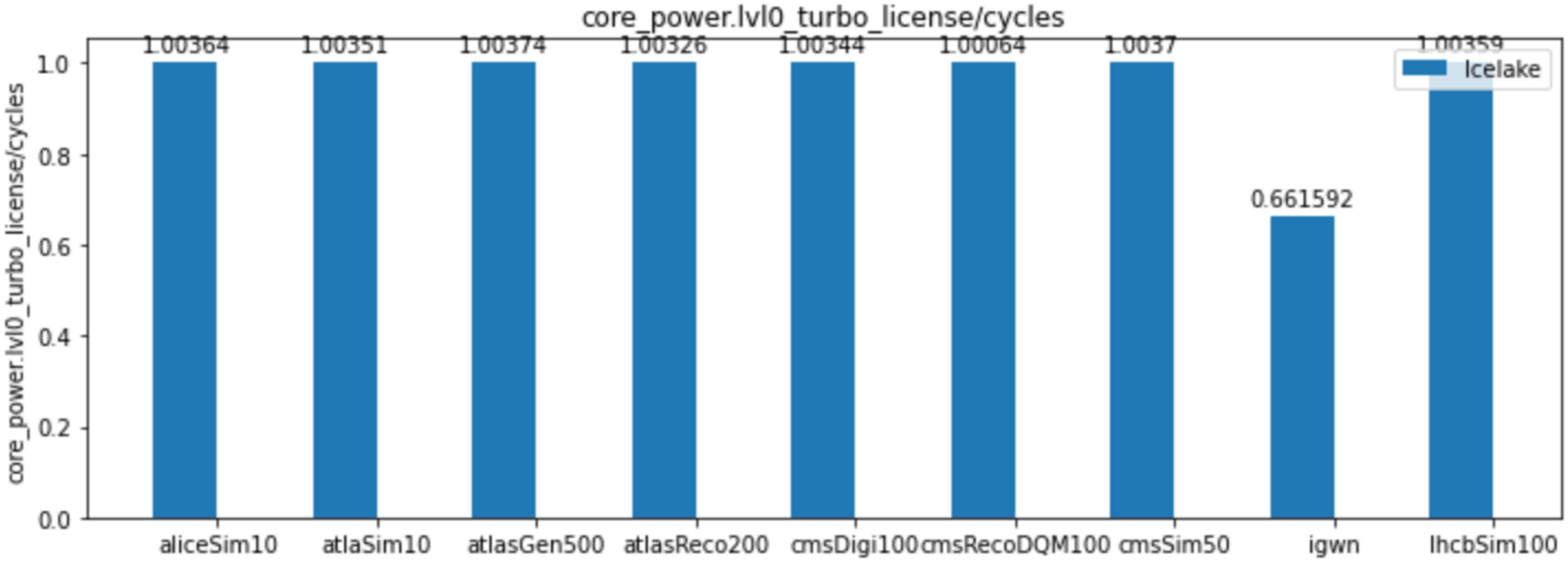
code compiled for SSE. Presence of AVX (even AVX512 for igwn) means that “fat libriaries” are used



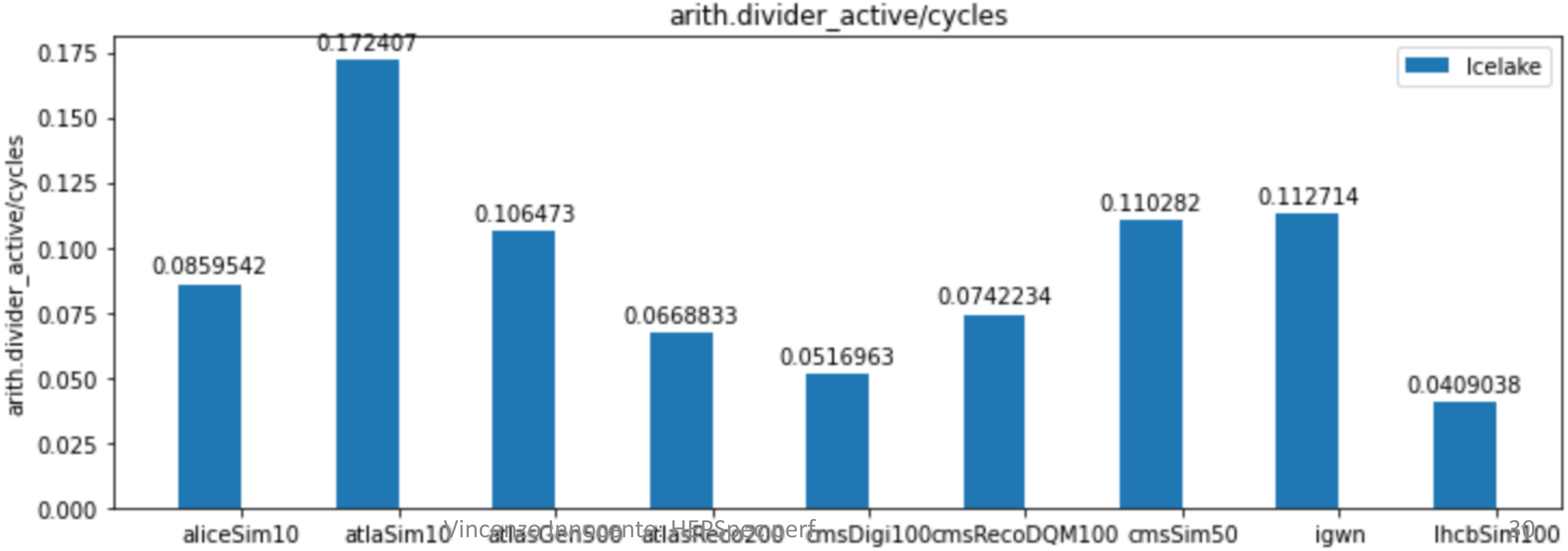




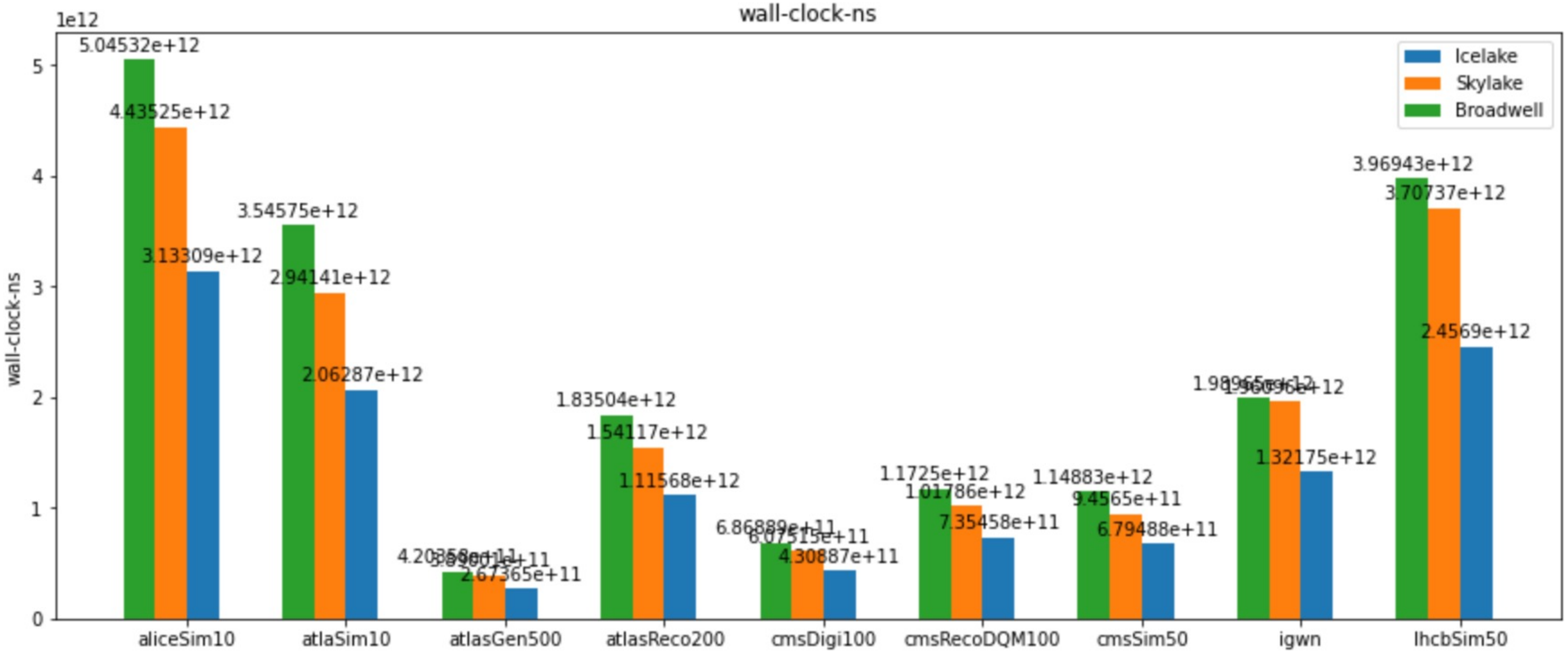
Freq throttling

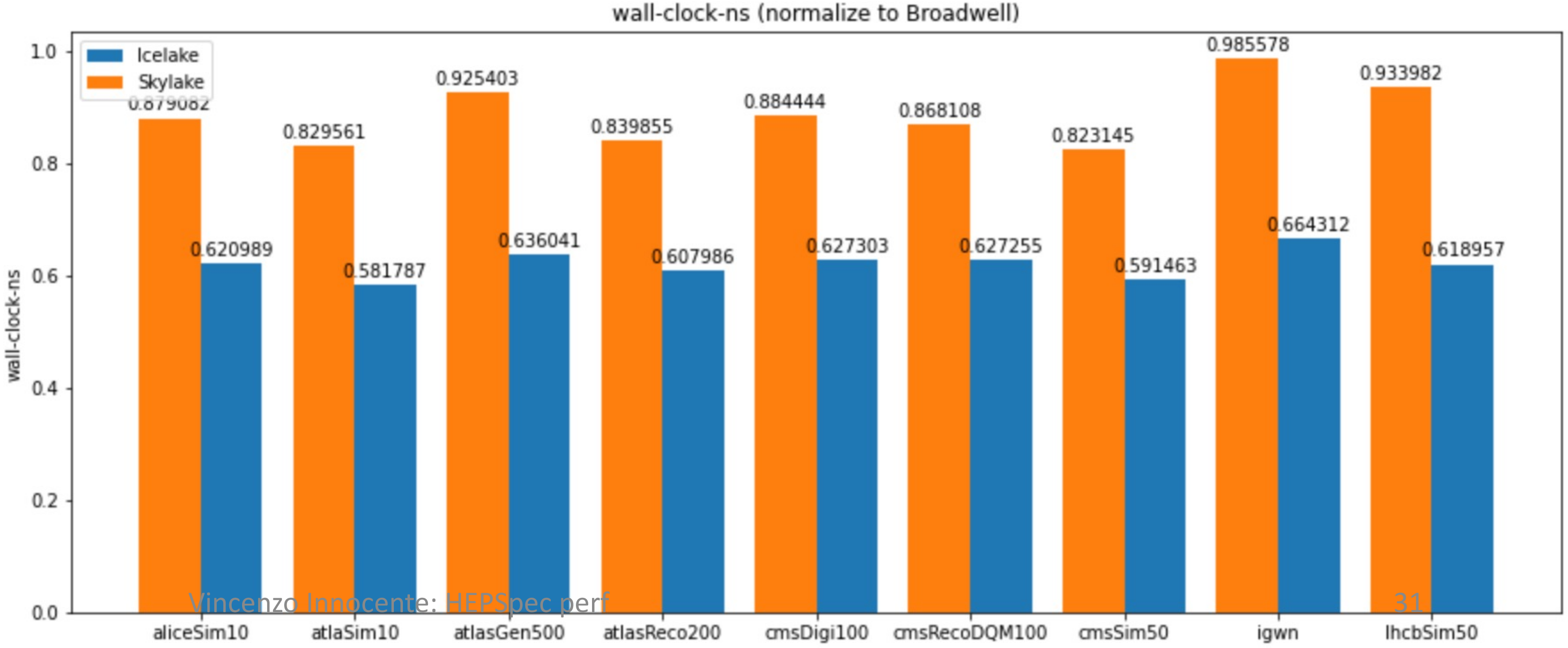


divisions and sqrt (fraction of total cycle)



Wall clock

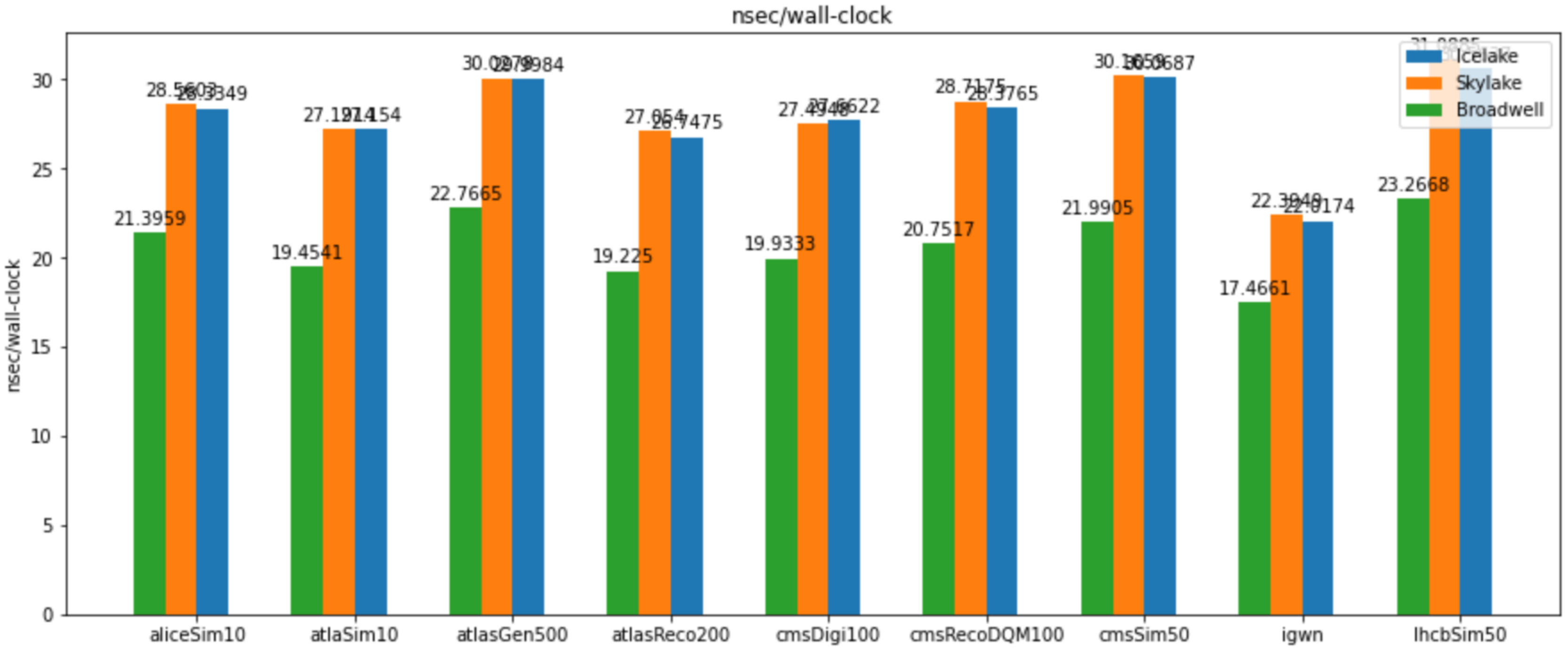


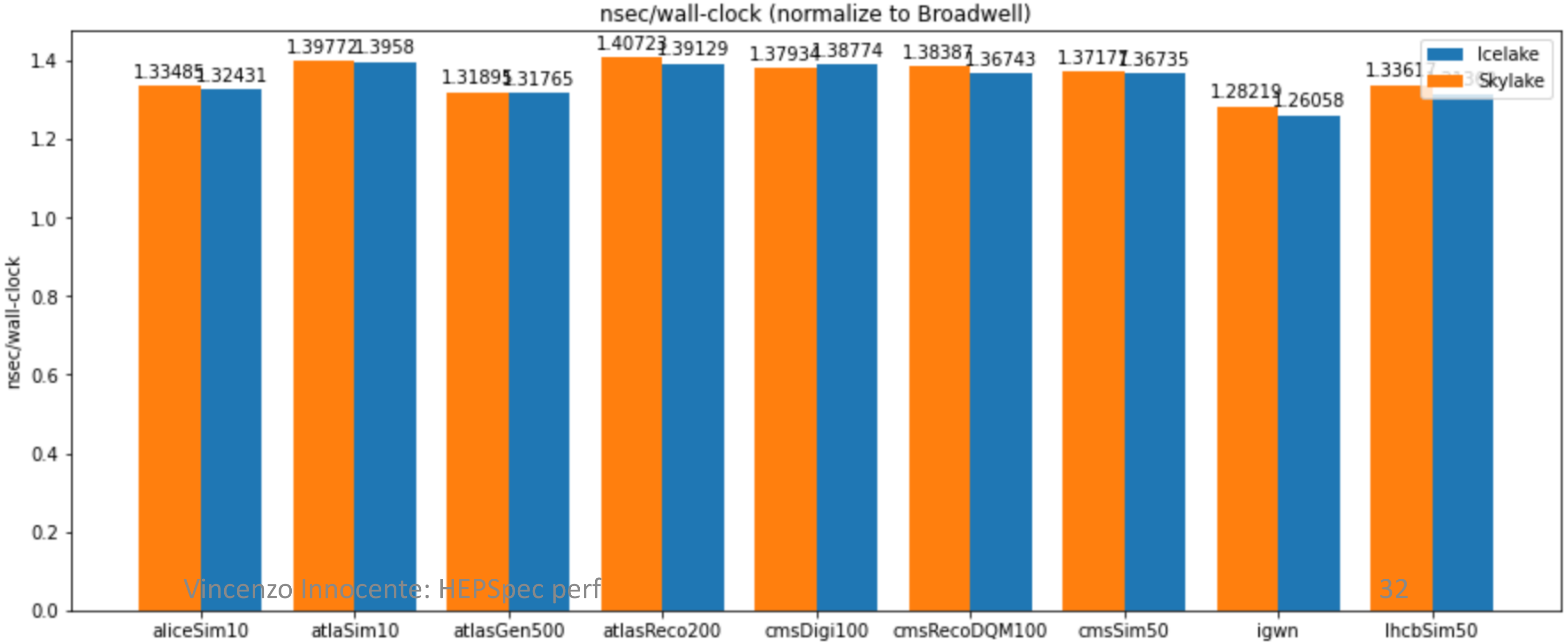


Icelake 40% faster than Broadwell

10->15% faster than Skylake

Thread efficiency: Task time / Wall clock (should be either 32 or 24)





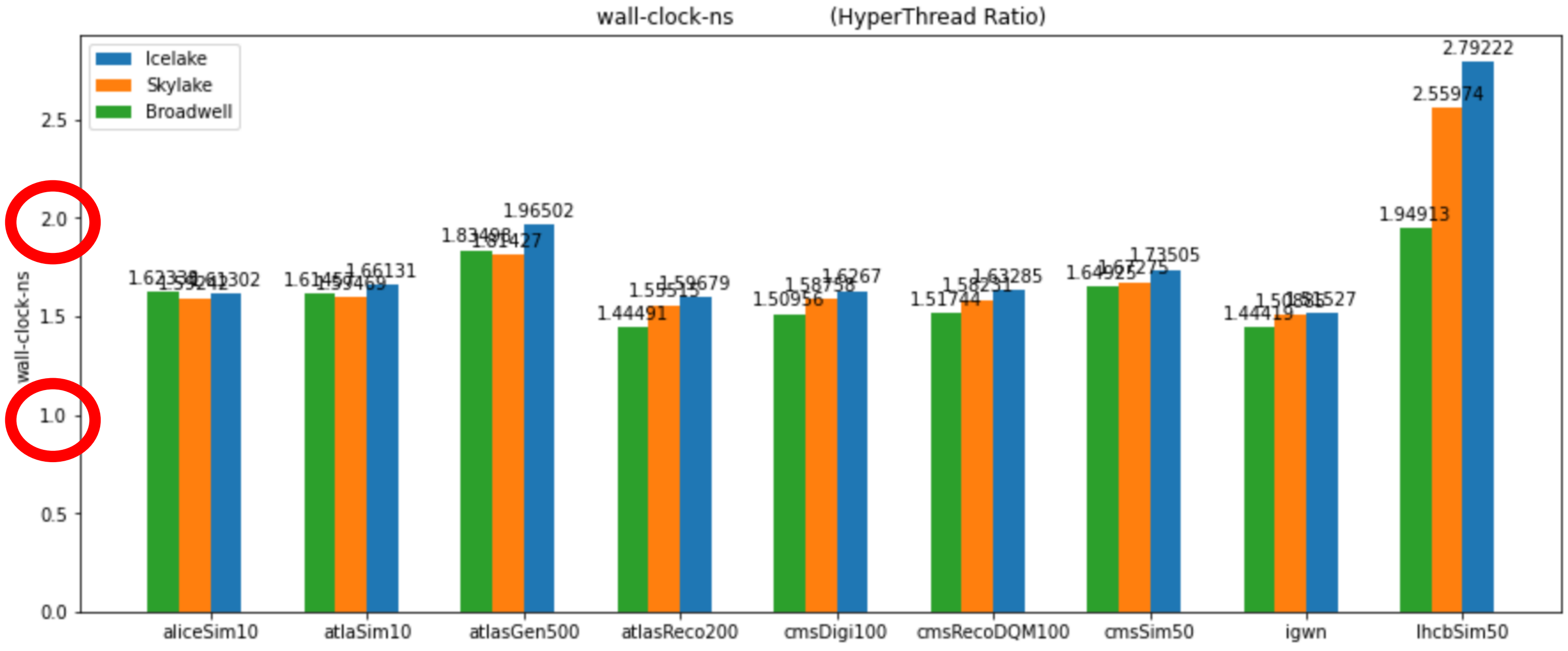
HyperThread efficiency:

Wall clock Ratio:

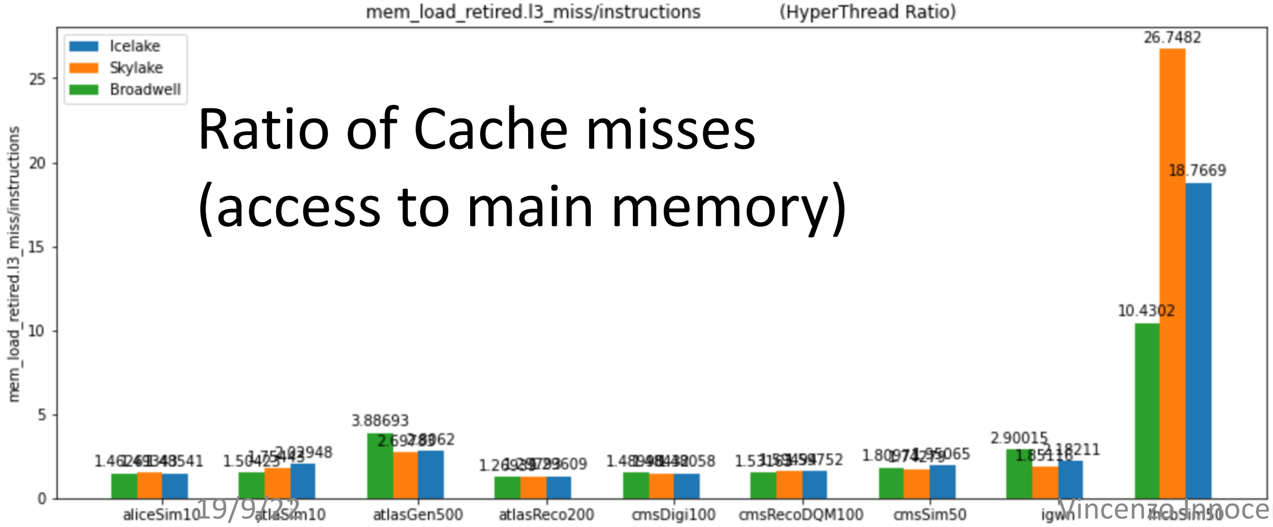
1: fully efficient

2: zero efficient

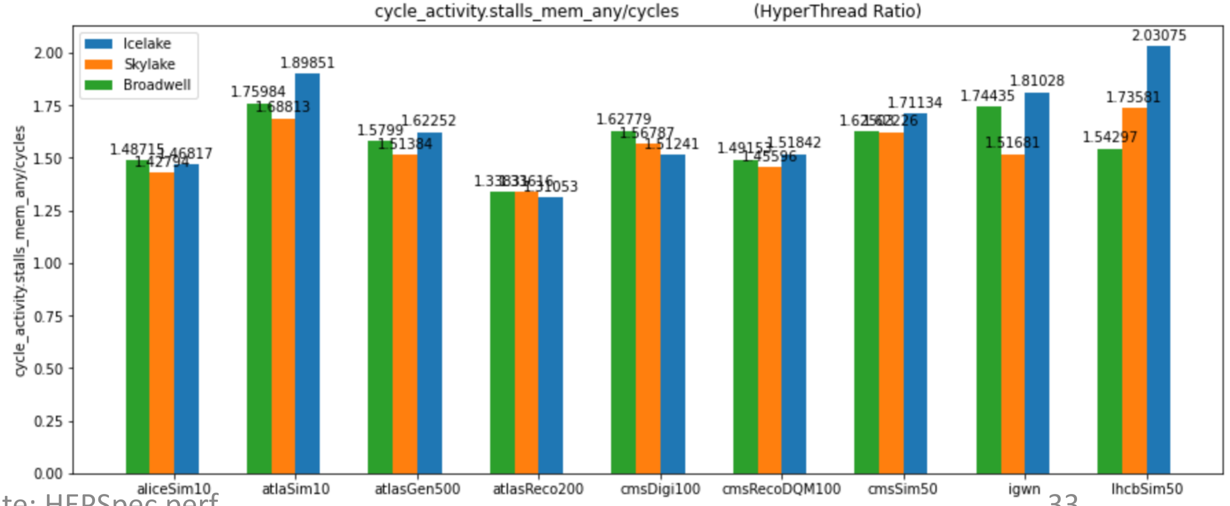
>2: penalizing



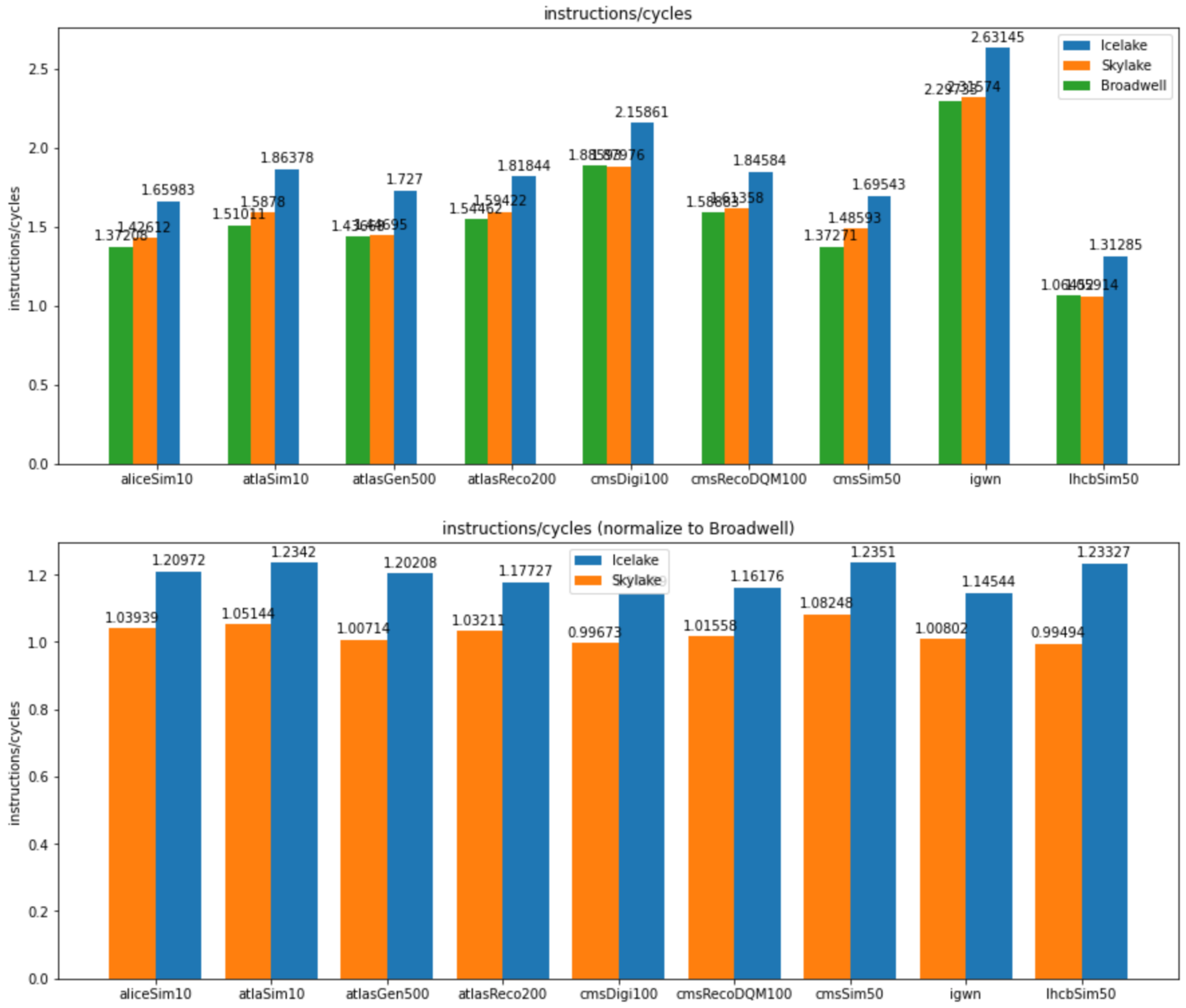
Ratio of Cache misses (access to main memory)



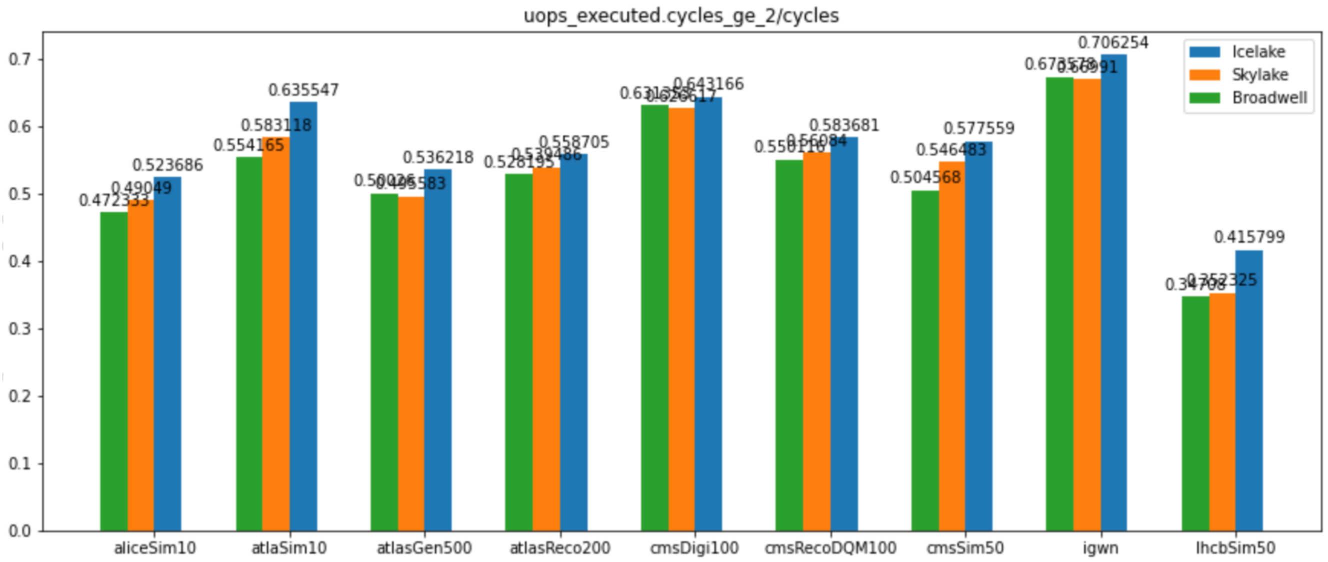
Total “cost” (in cycles of memory access)

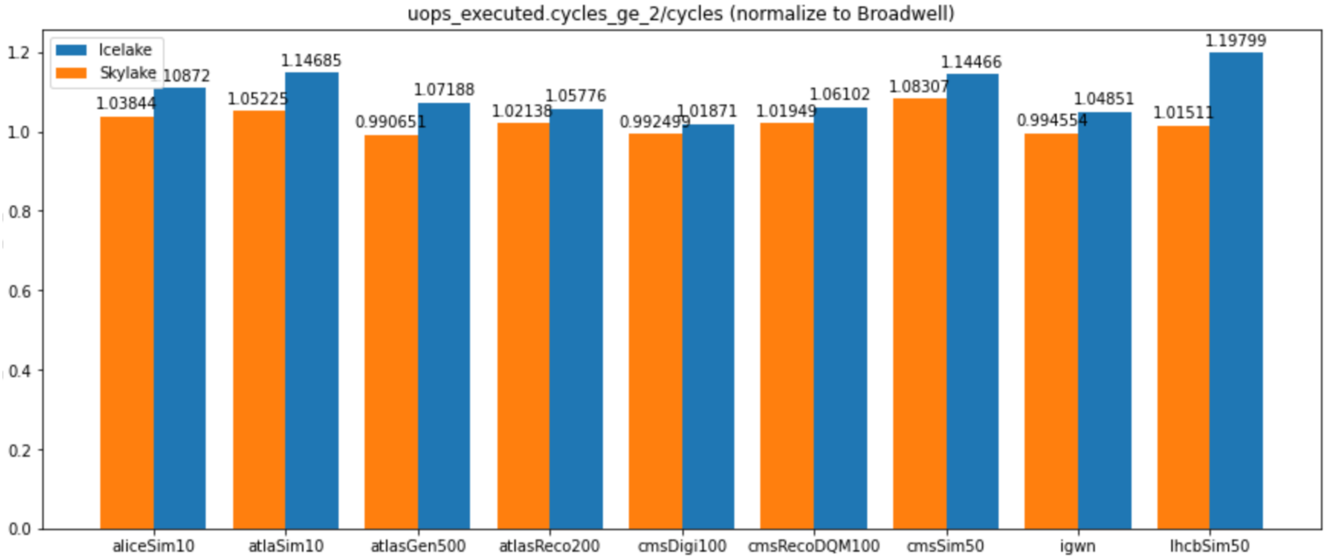


CPU efficiency IPC (max 4)

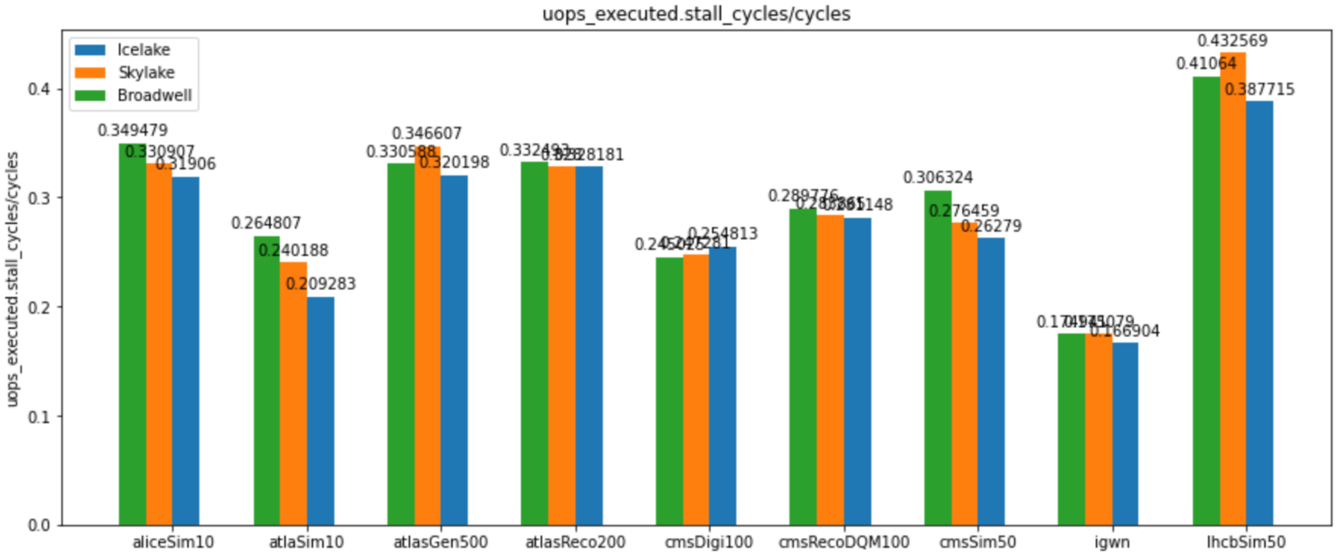


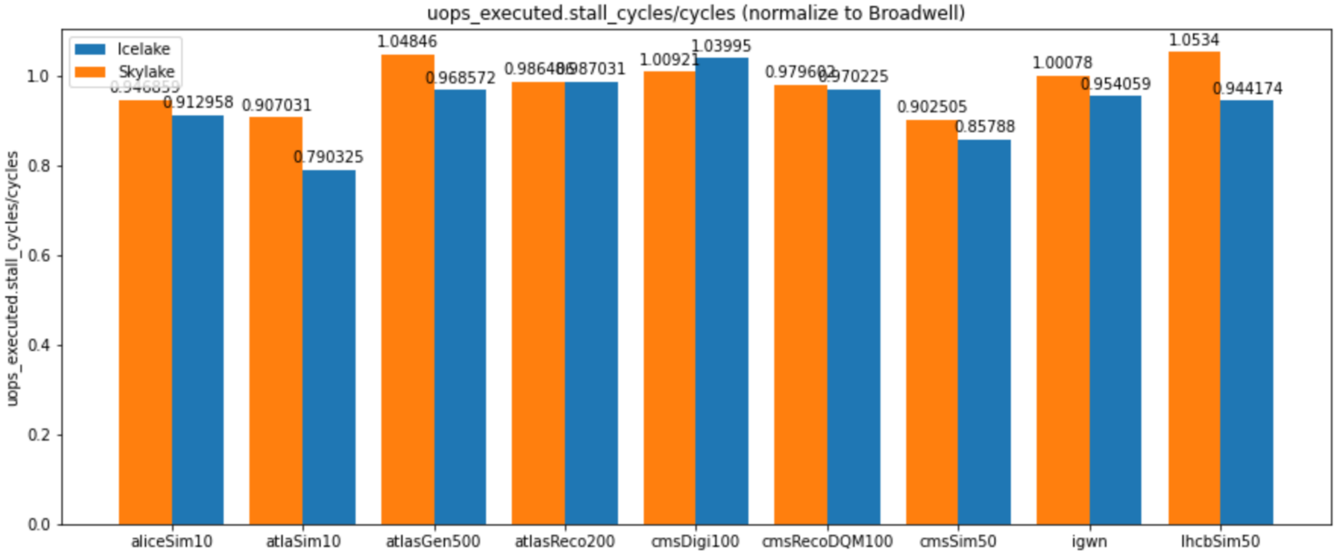
Parallelism: #cycles >1 instrutions (2 or more)





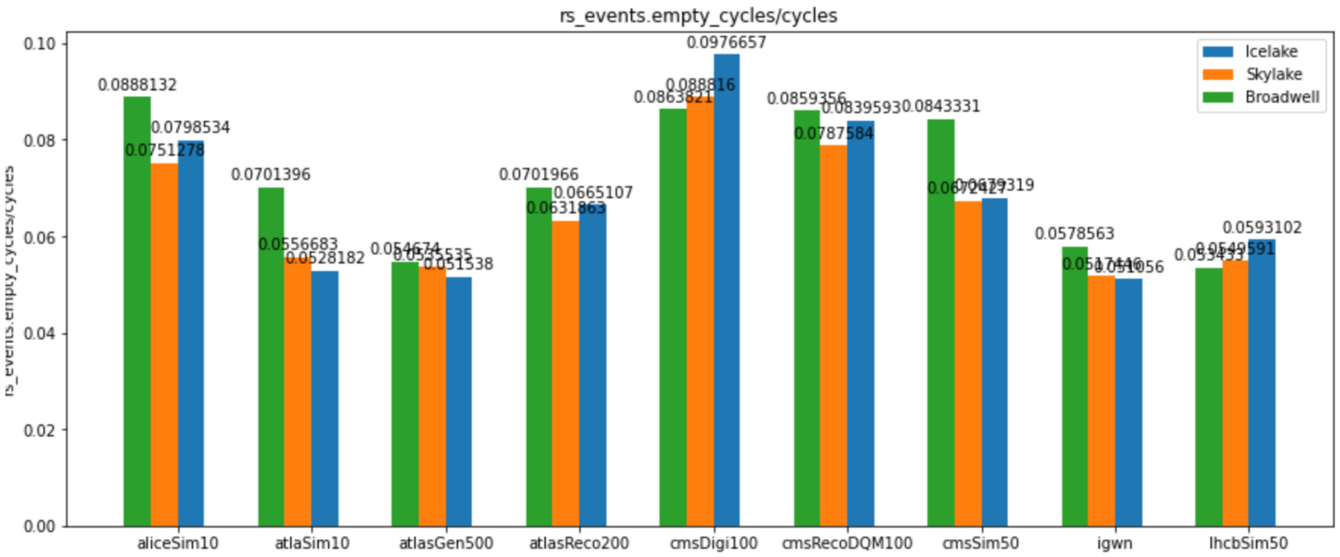
Stalls cycles where no instruction executed

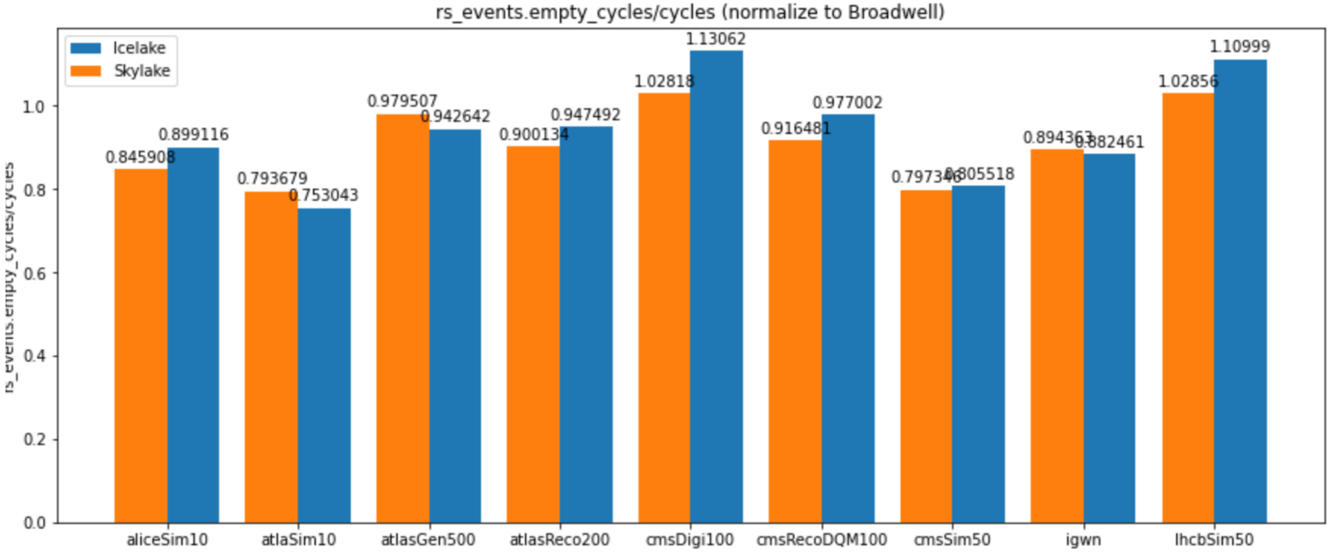




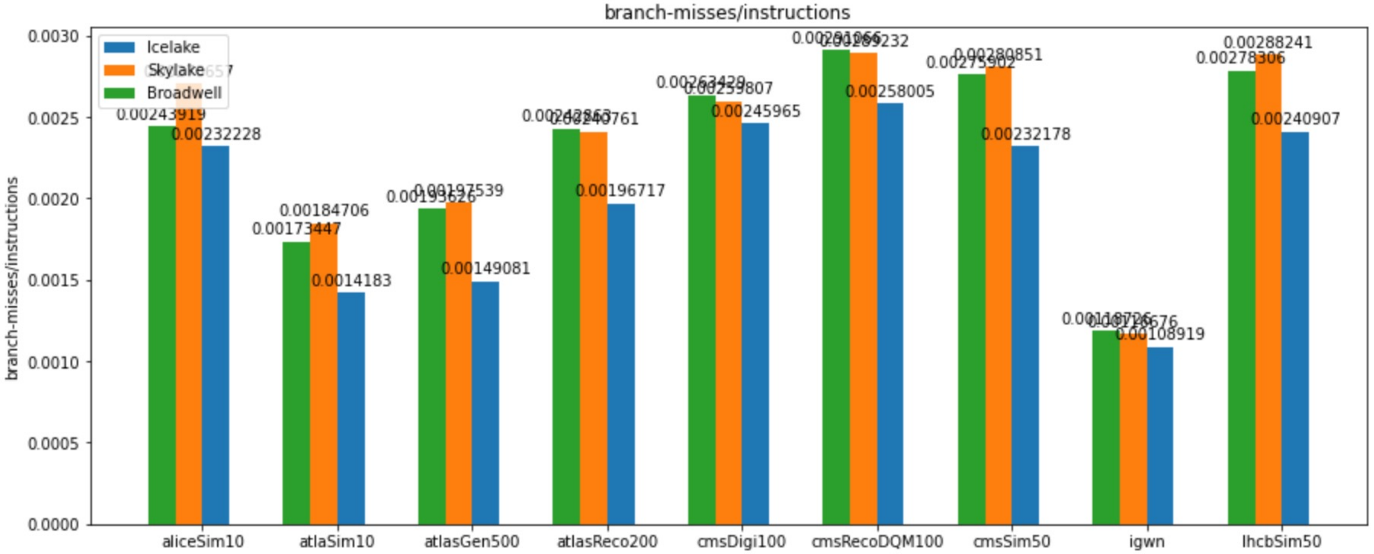
CPU starvation

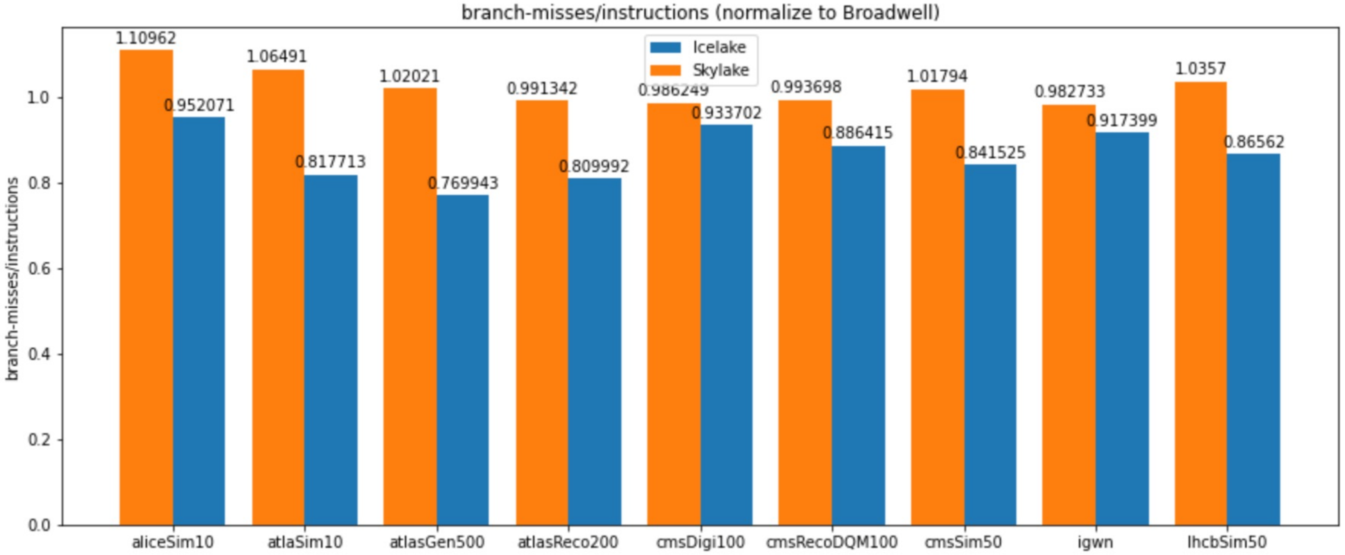
empty reserve-station





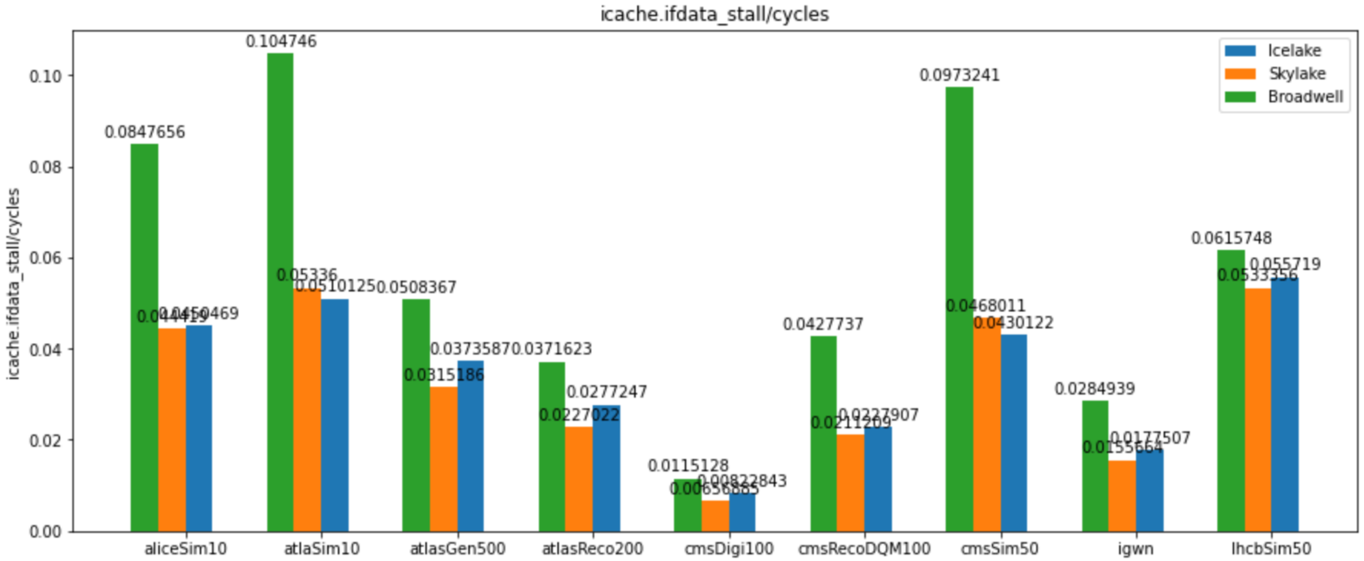
Branch misses (bad speculation) cost ~20 cycles

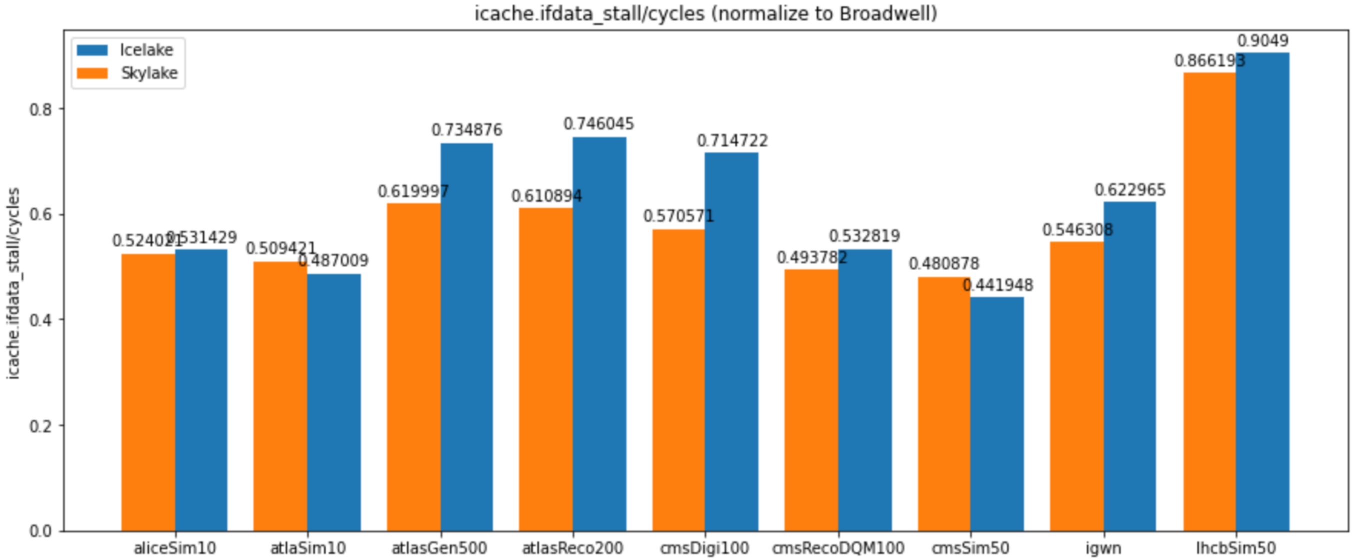




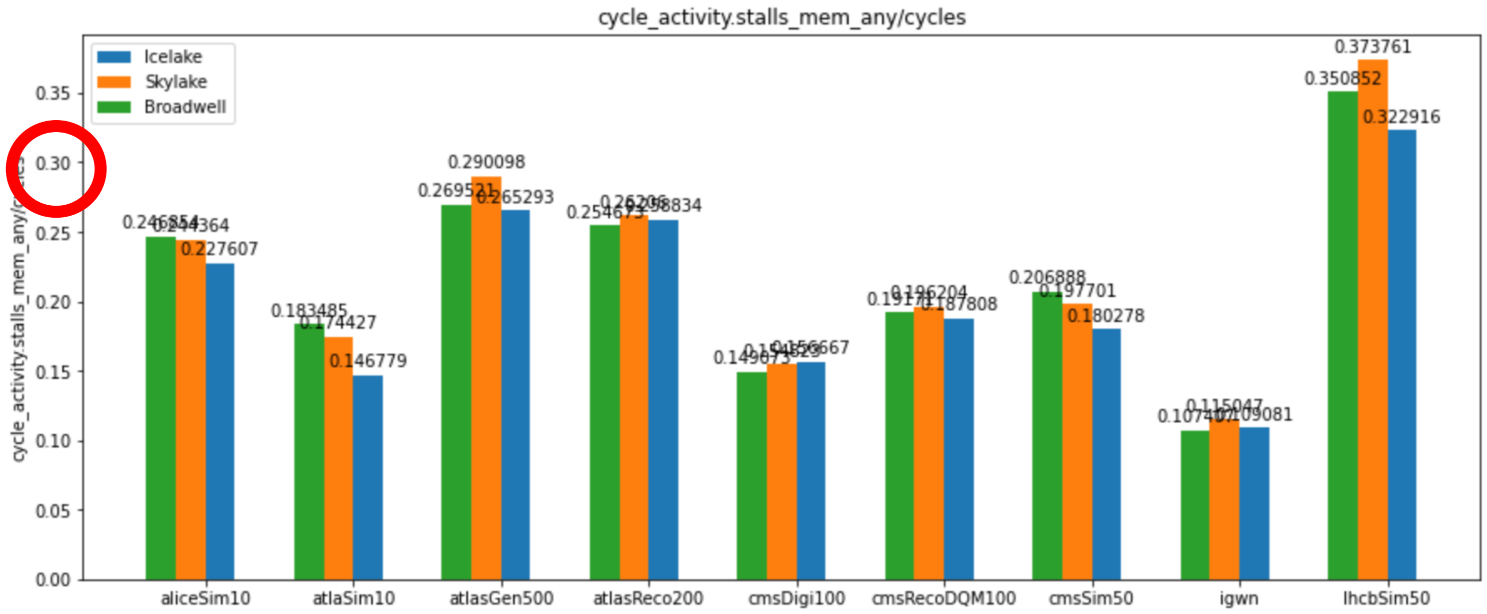
https://users.elis.ugent.be/~leeckhou/papers/ispass06-eyerman.pdf

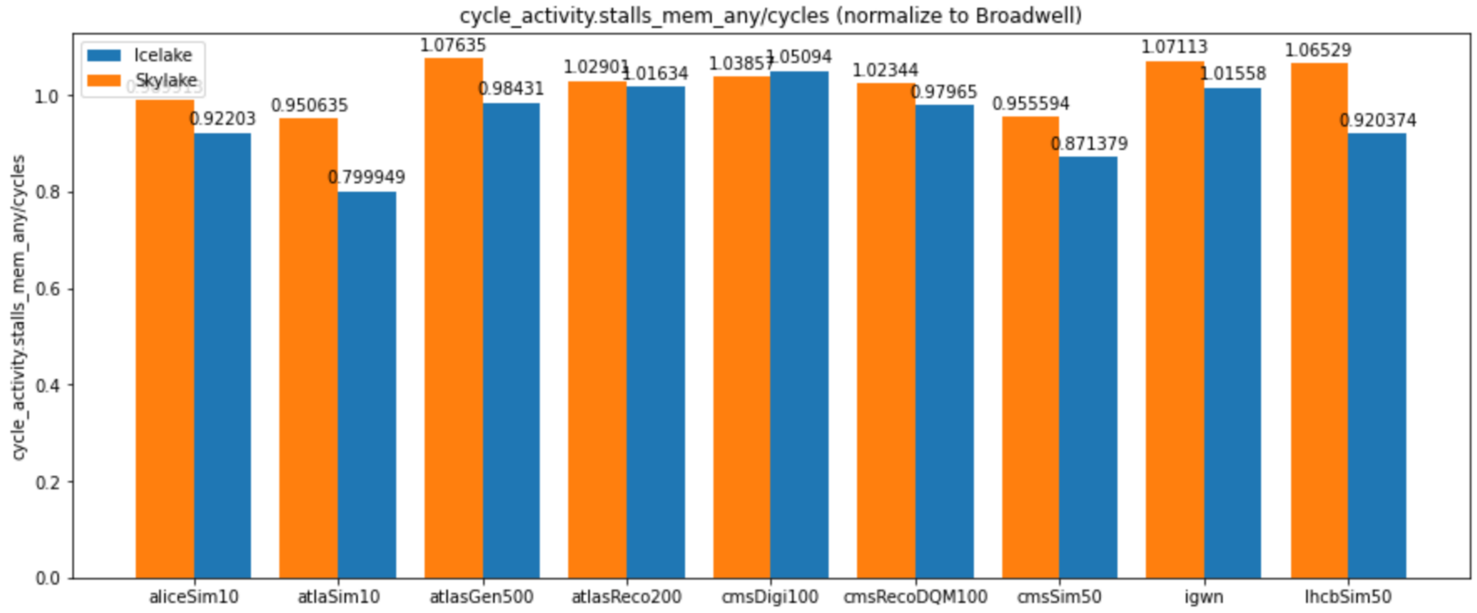
Instruction cache stall



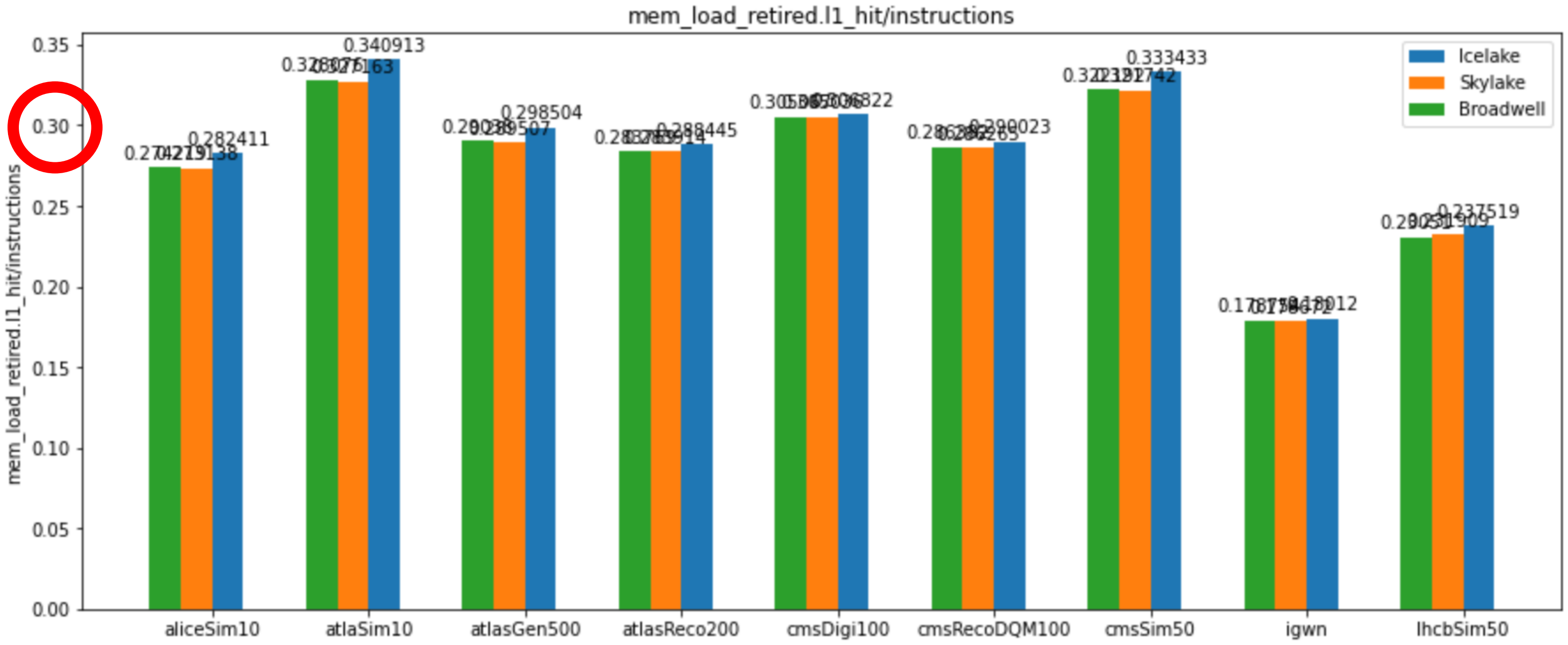


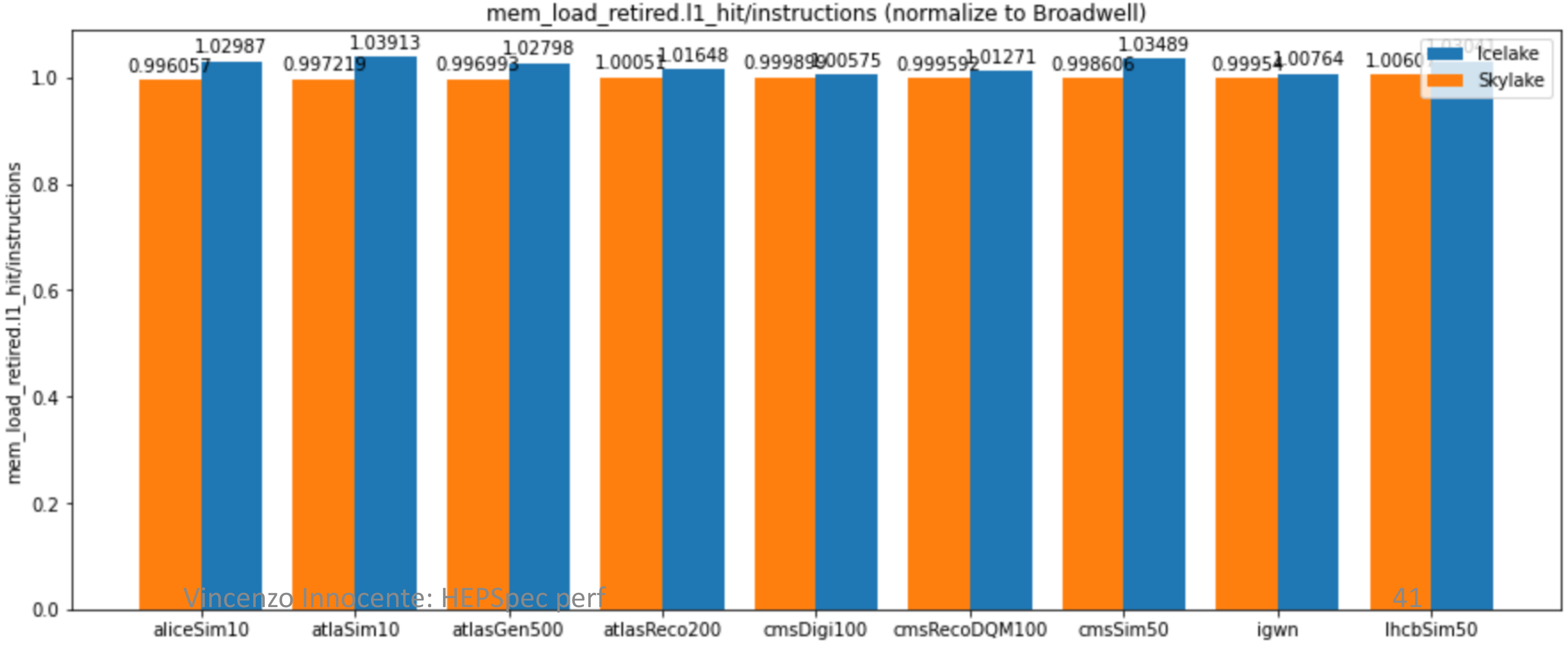
Memory stalls



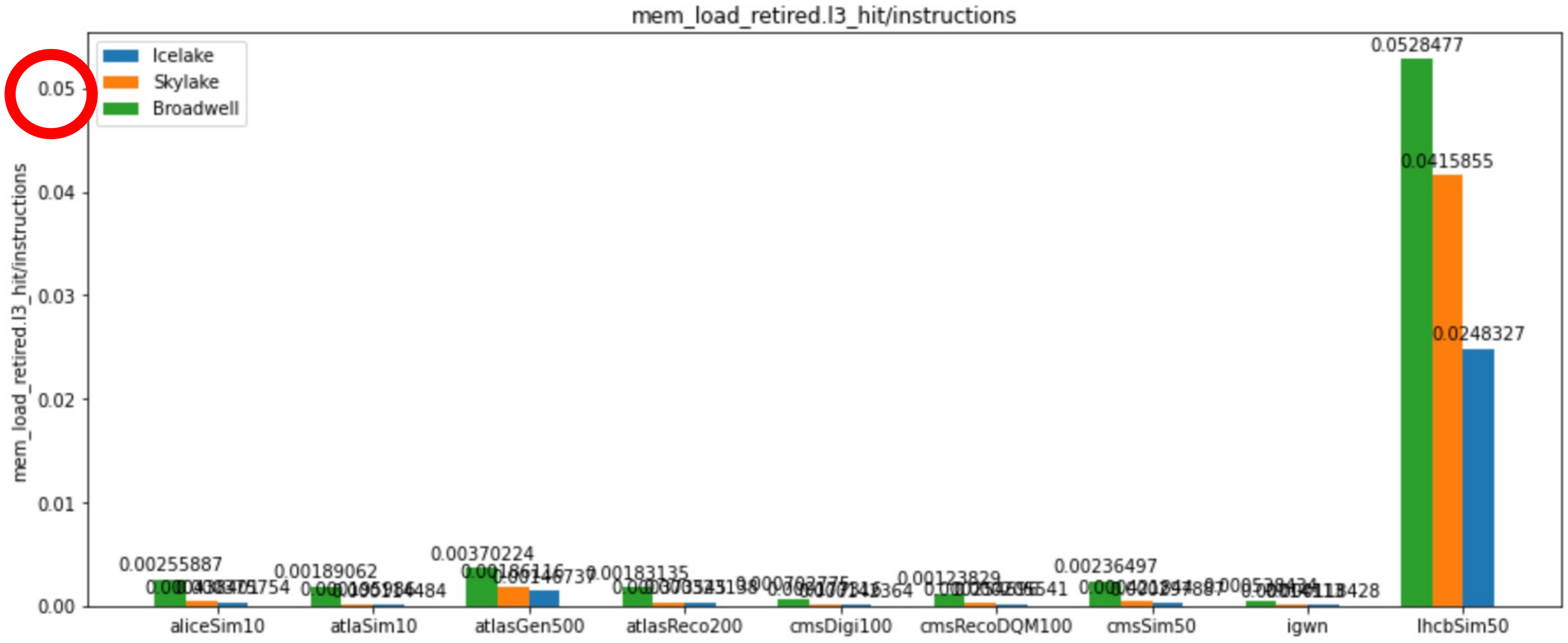


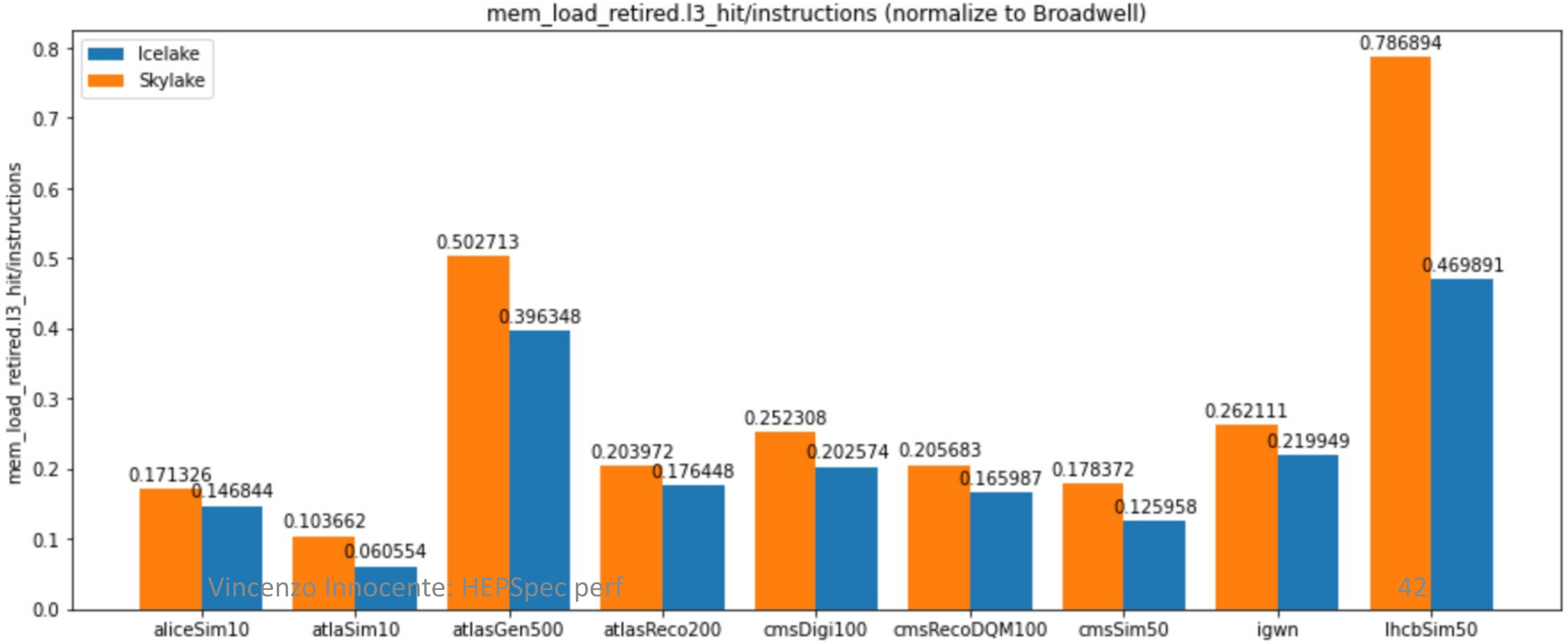
L1 cache access % total instruction (4 cycles latency)



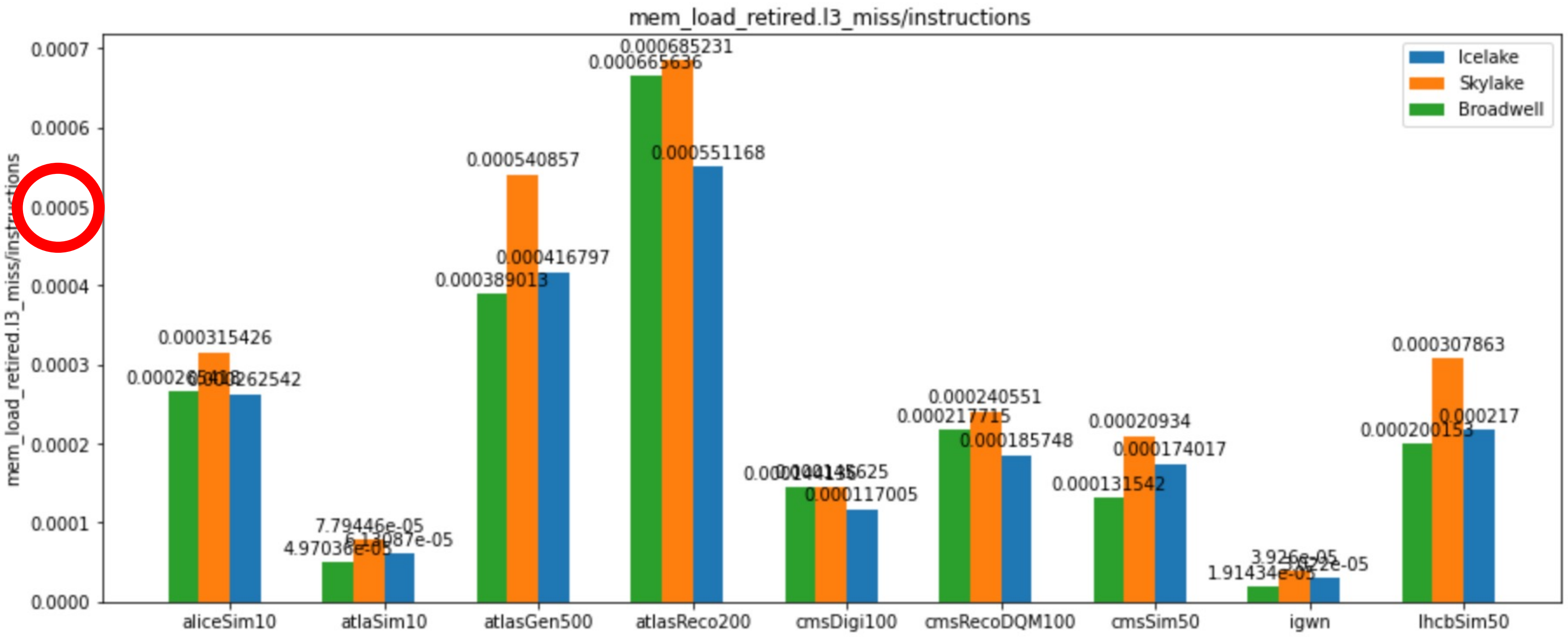


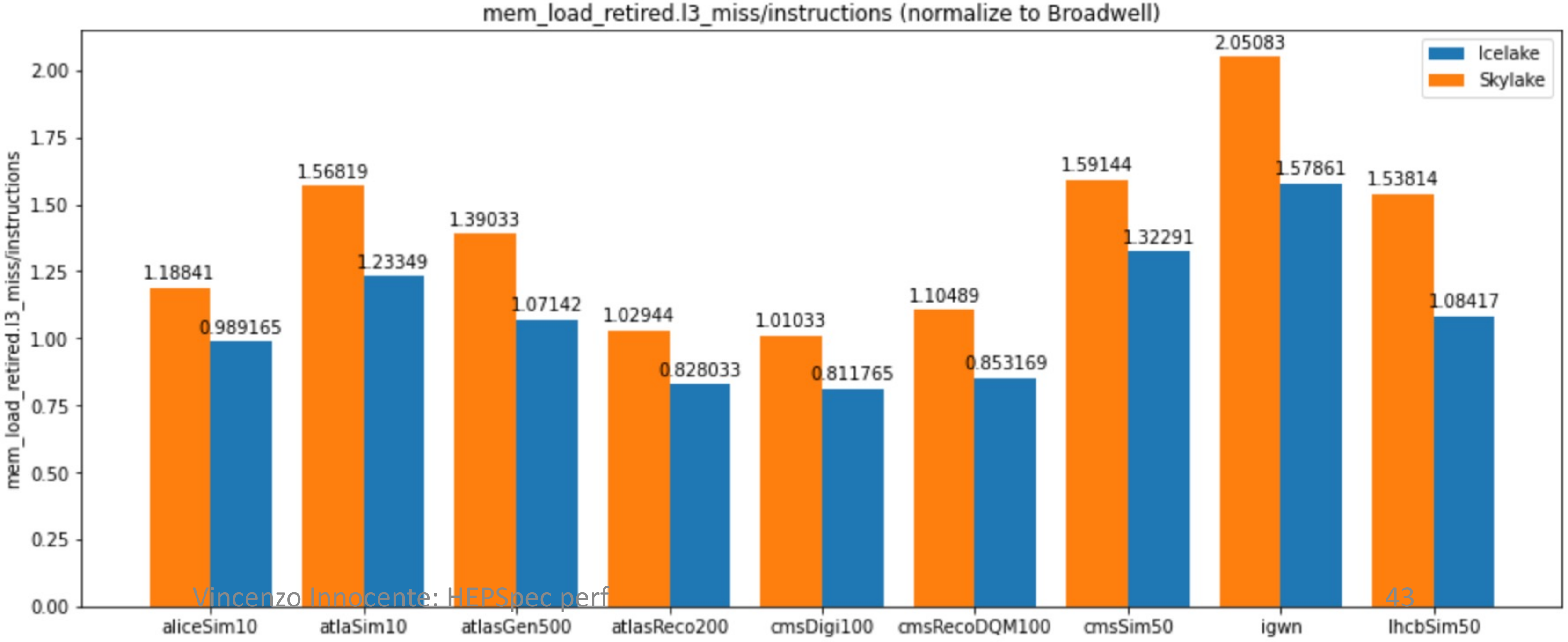
L3 cache access (50 cycles latency)



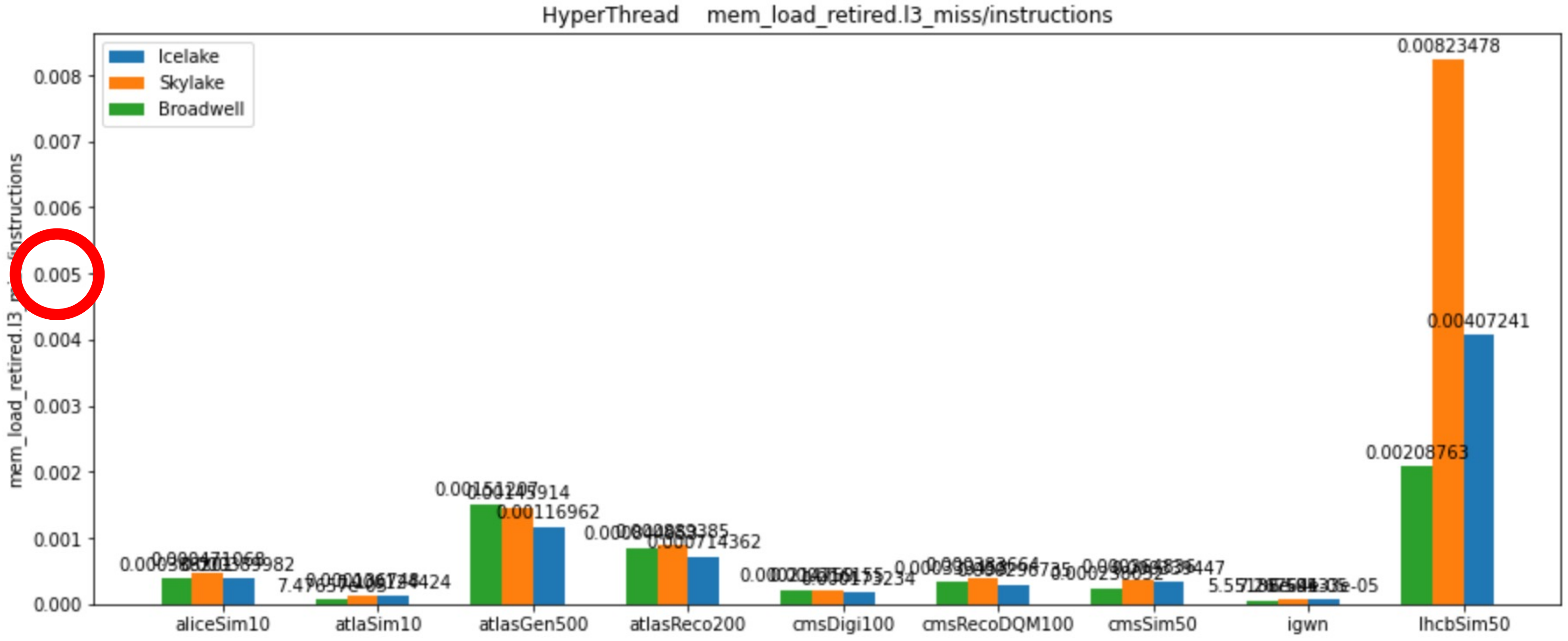


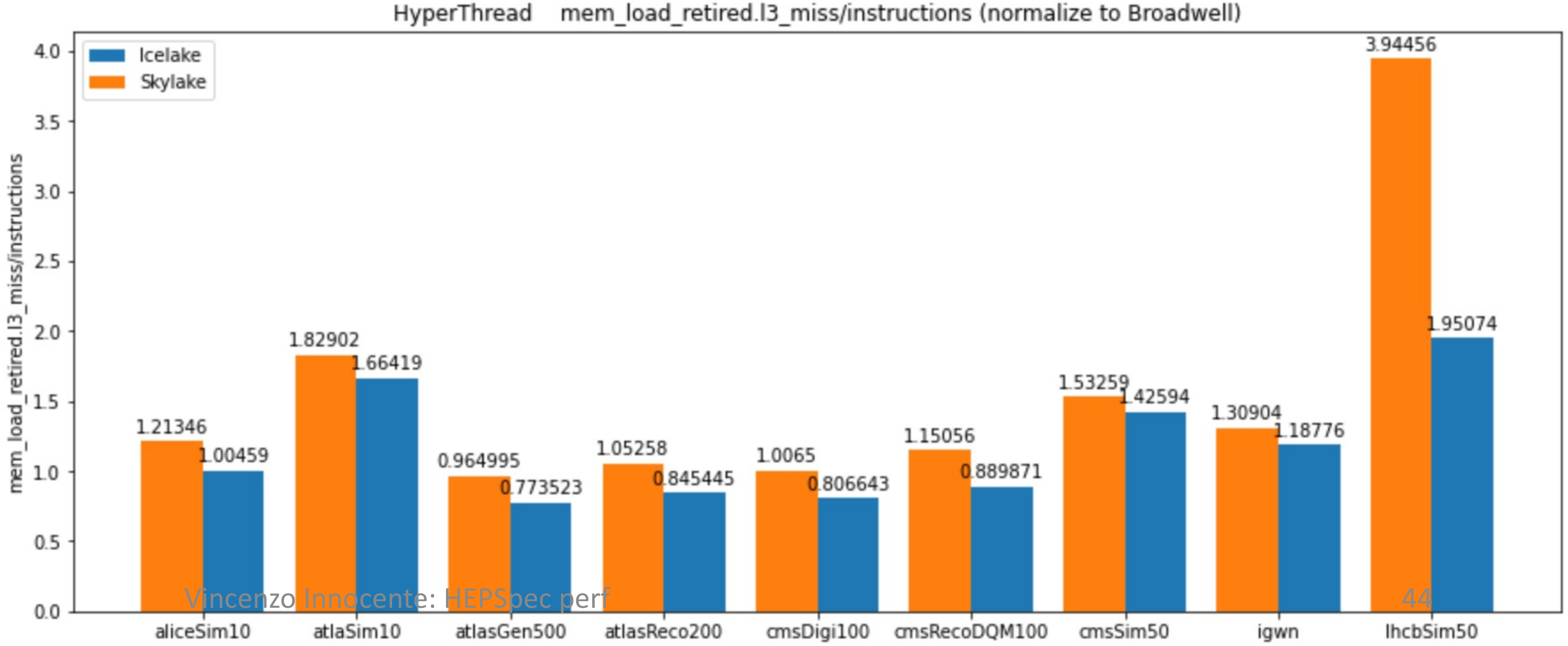
Main memory access (>200 cycles ~100 ns latency)





HyperThread Main memory access (>200 cycles ~100 ns latency)





Summary (3)

- Icelake is faster and “wider” than previous Intel models
  - In general all metrics improve (not as much as advertised)
  - Resource hungry wf will profit more
  - Higher memory access (cache misses) w/r/t Broadwell
- HyperThread efficiency is limited by memory access
  - On Icelake is in general less performant
  - For LHCb simulation is even penalizing
- 0.2% of avx512 instructions are costing a 10% frequency reduction on Skylake (running 33% of the time al lower frequency)

Conclusions

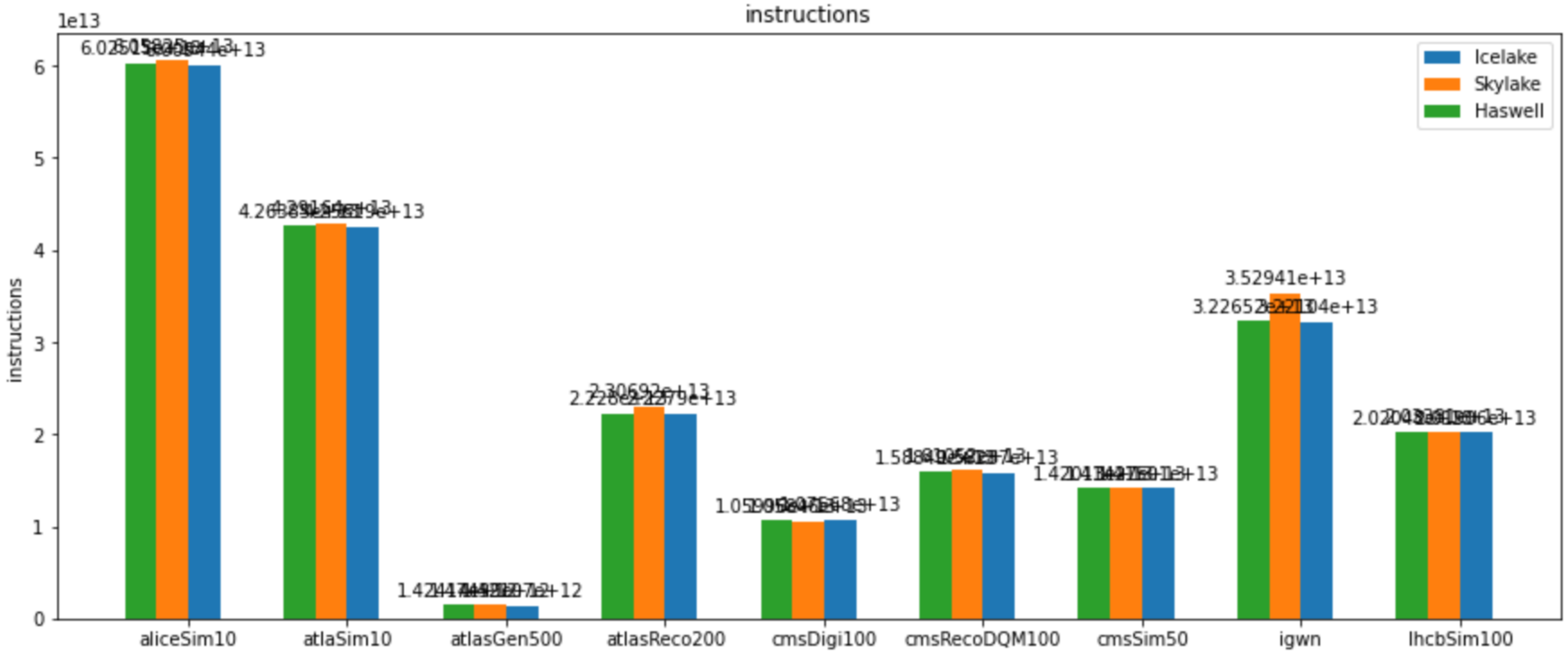
- I have shown how simple, light tools (turbostat, perf) can provide critical insight on the performance of workflows and detect:
  - Hotspots
  - Resource hungriness
  - Anomaly in Metrics
- I will advice to add those tools to the standard singularity benchmark instance. This will allow
  - To produce detailed *“perf report”* w/o access to library on metal
  - To generate reports by *“turbostat”* and *“perf stat”* at each benchmark run

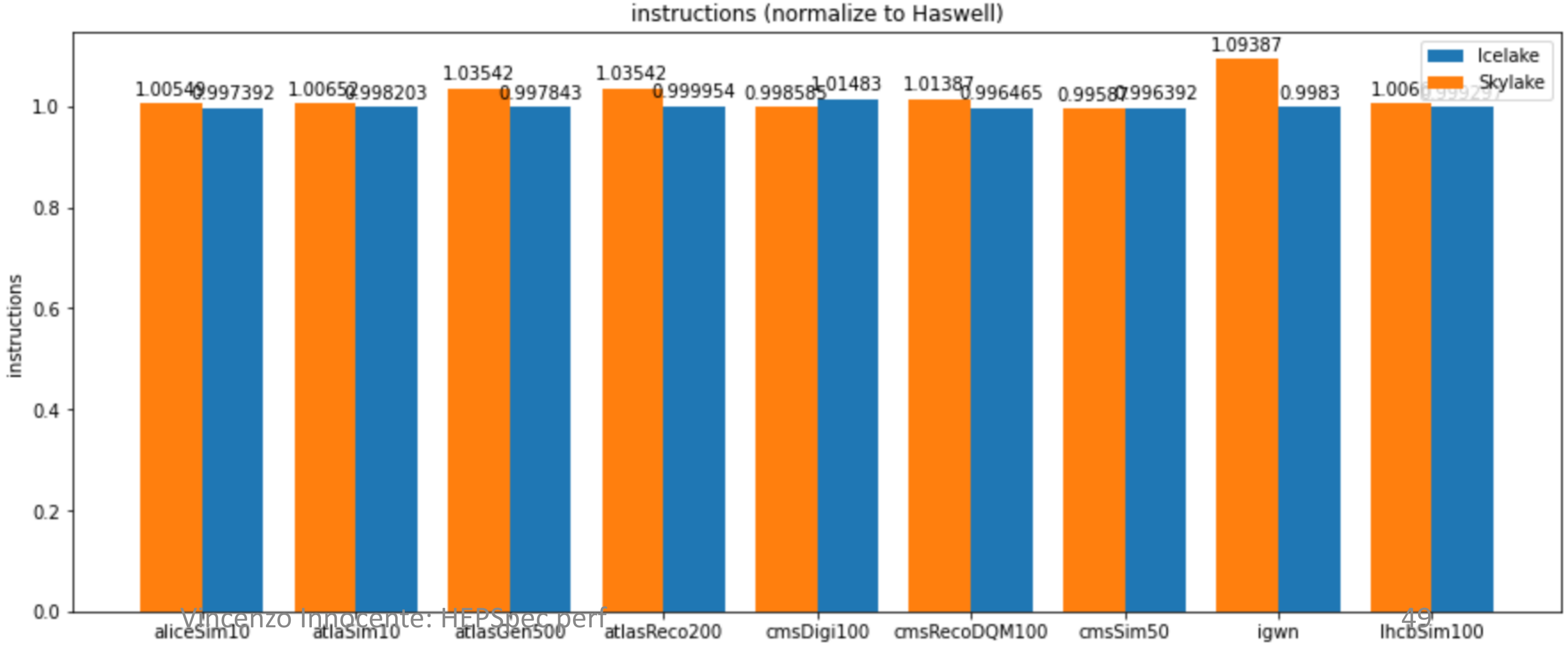
Backup

Detailed Perf Statistics

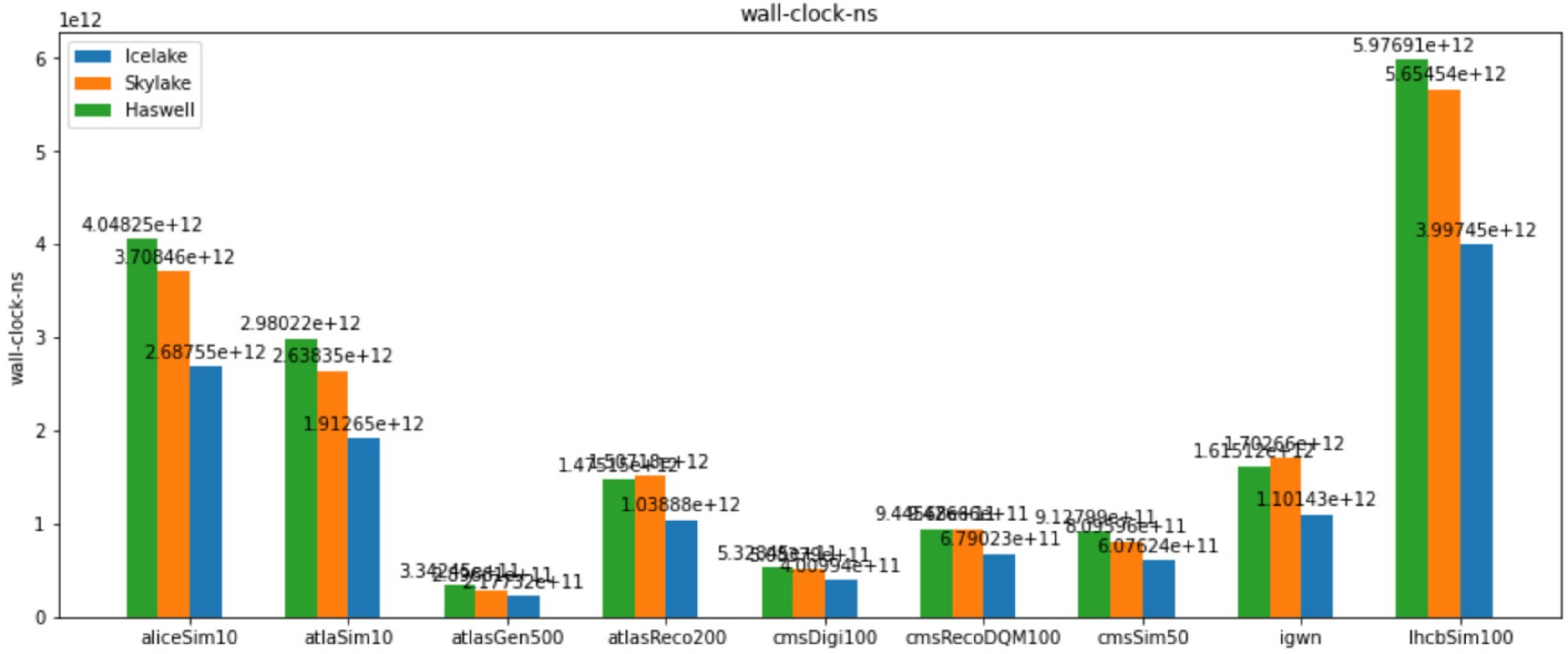
Single Process

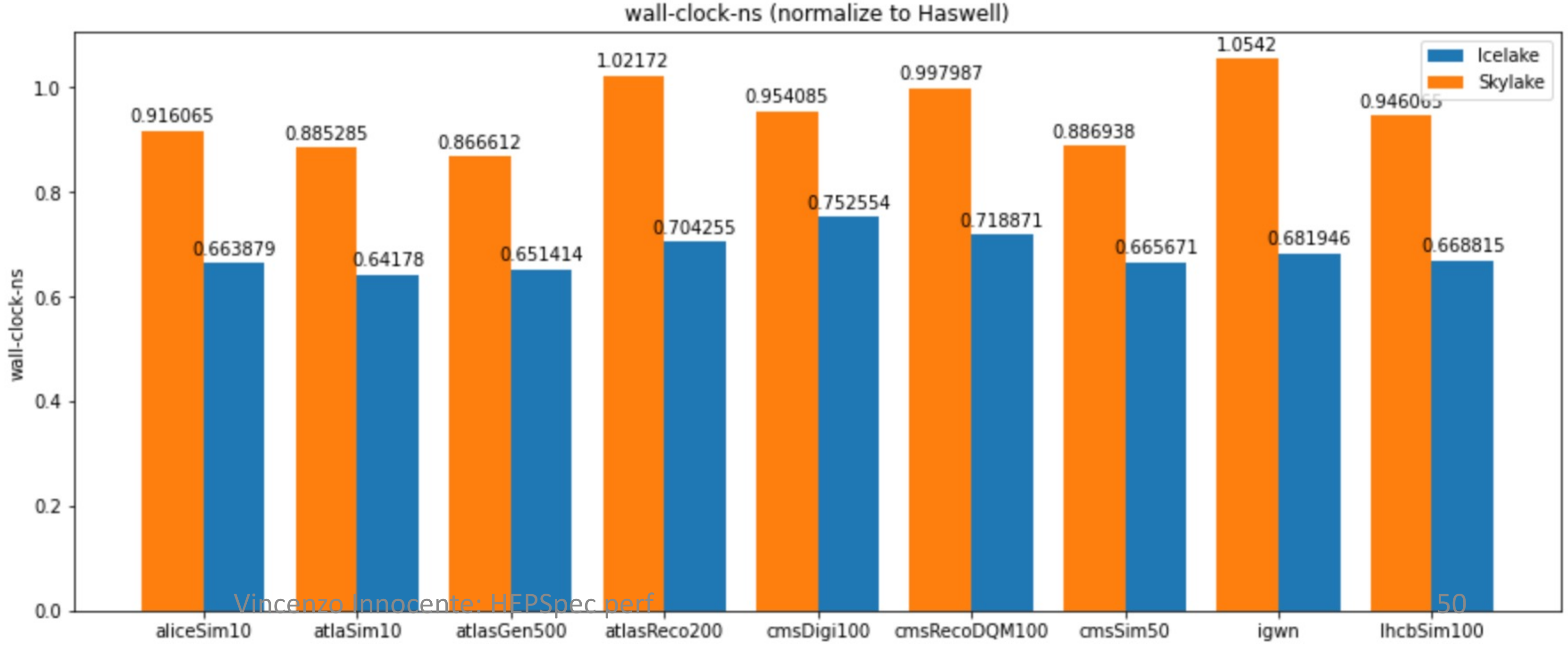
#Instructions good: same program!



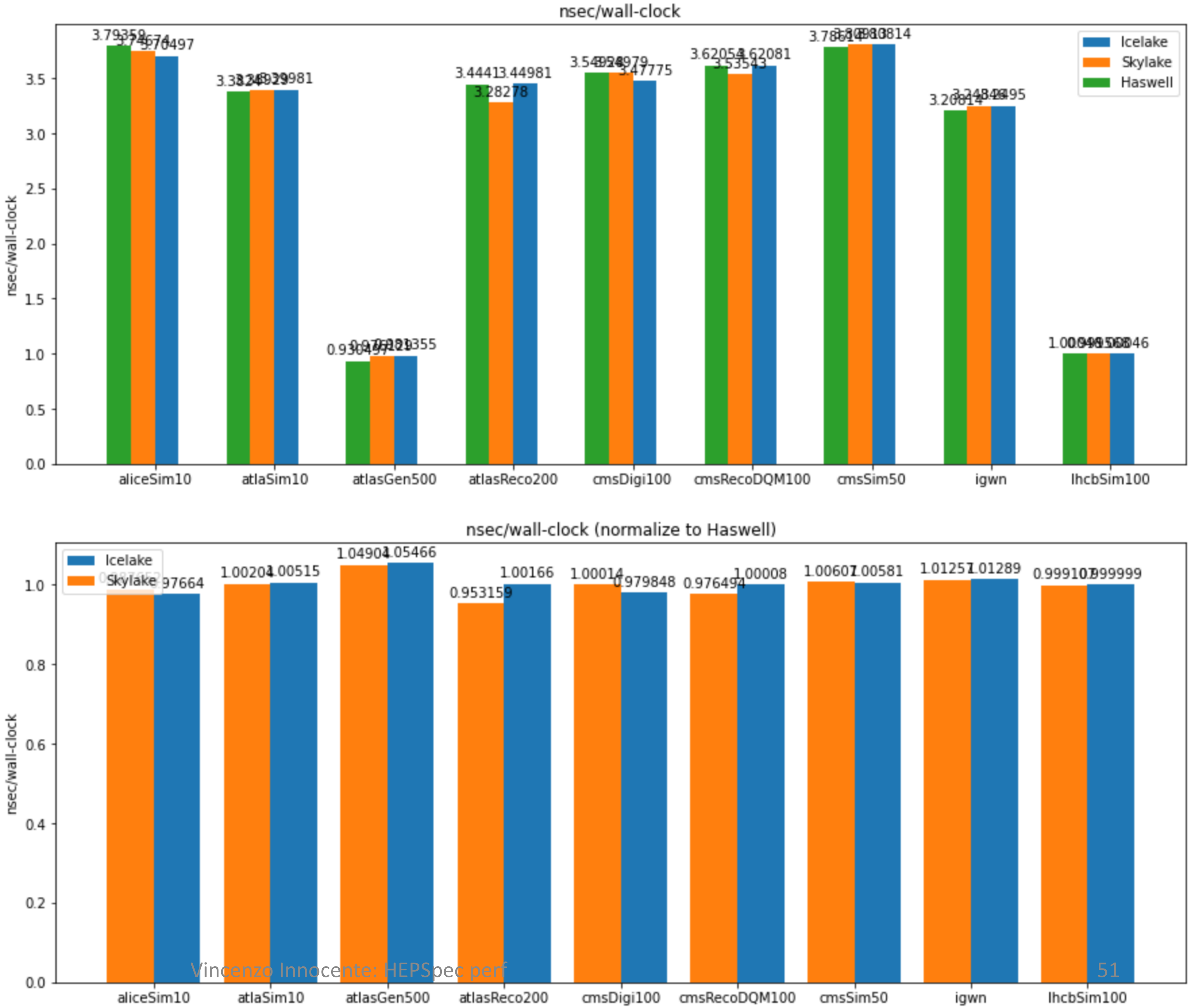


Wall clock

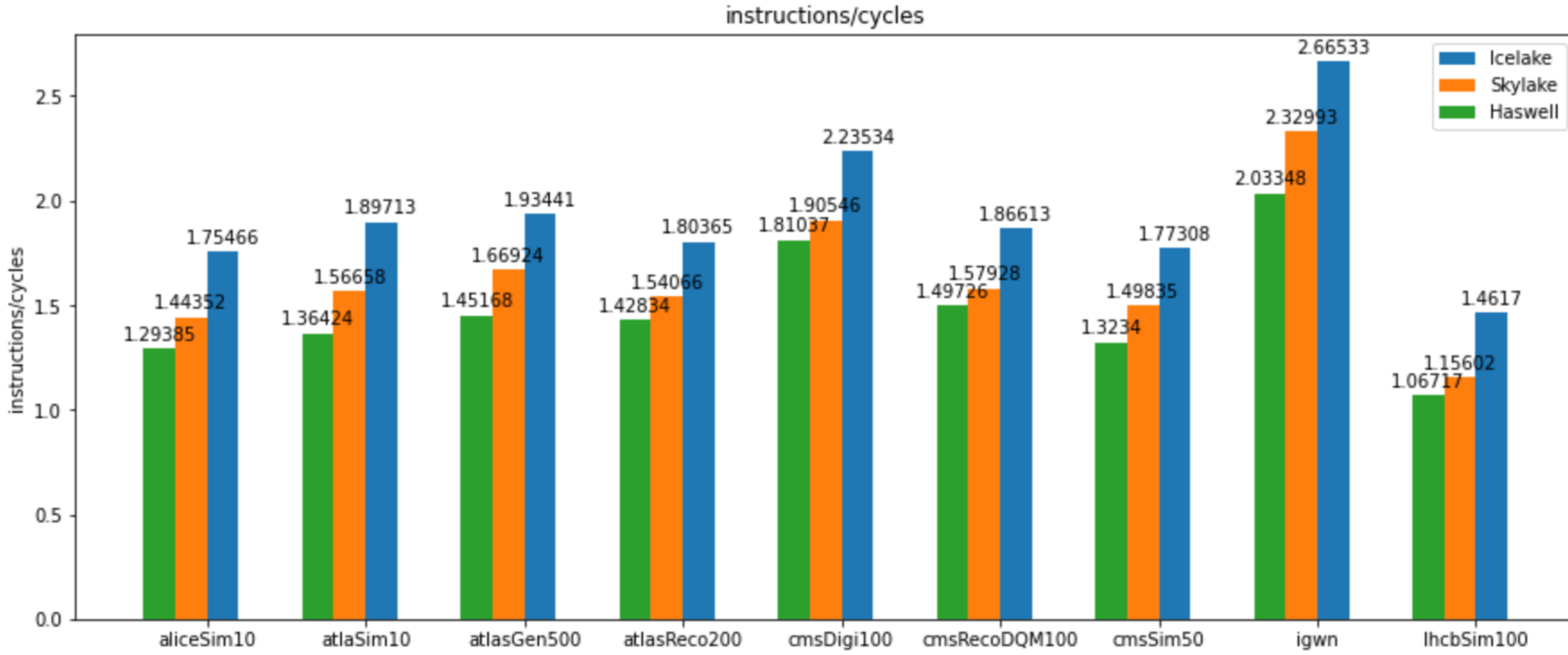

wall-clock-ns
1e12
Icelake
Skylake
Haswell
wall-clock-ns
0
1
2
3
4
5
6
4.04825e+12
3.70846e+12
2.68755e+12
2.98022e+12
2.63835e+12
1.91265e+12
3.34246e+11
2.17732e+11
1.47513e+12
1.50718e+12
1.03888e+12
4.00994e+11
6.79023e+11
9.12799e+11
8.09596e+11
6.07624e+11
1.61512e+12
1.70266e+12
1.10143e+12
5.97691e+12
5.65454e+12
3.99745e+12
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim50
igwn
lhcbSim100

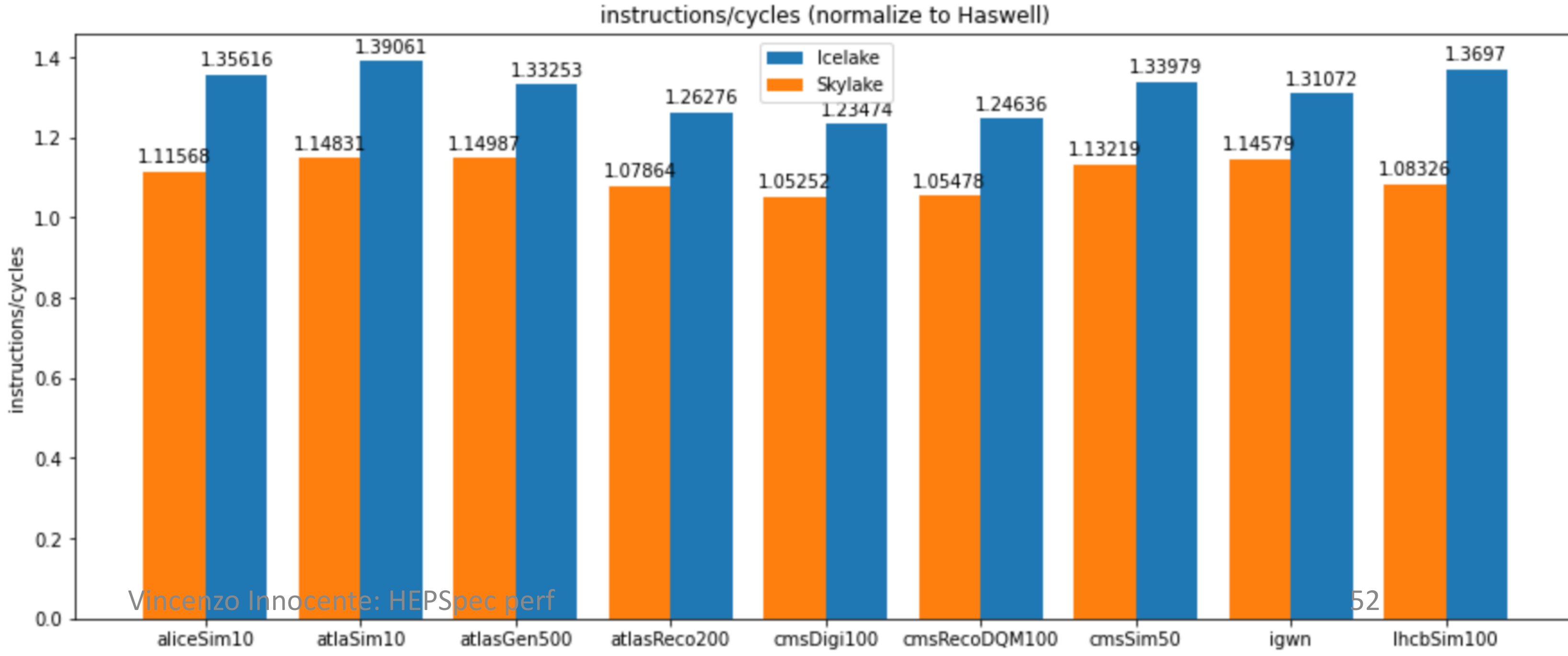

wall-clock-ns (normalize to Haswell)
Icelake
Skylake
wall-clock-ns
0.0
0.2
0.4
0.6
0.8
1.0
0.916065
0.663879
0.885285
0.64178
0.866612
0.651414
1.02172
0.704255
0.954085
0.752554
0.997987
0.718871
0.886938
0.665671
1.0542
0.681946
0.946065
0.668815
aliceSim10
atlaSim10
atlasGen500
atlasReco200
cmsDigi100
cmsRecoDQM100
cmsSim50
igwn
lhcbSim100

Thread efficiency: Task time / Wall clock (should be either 4 or 1)



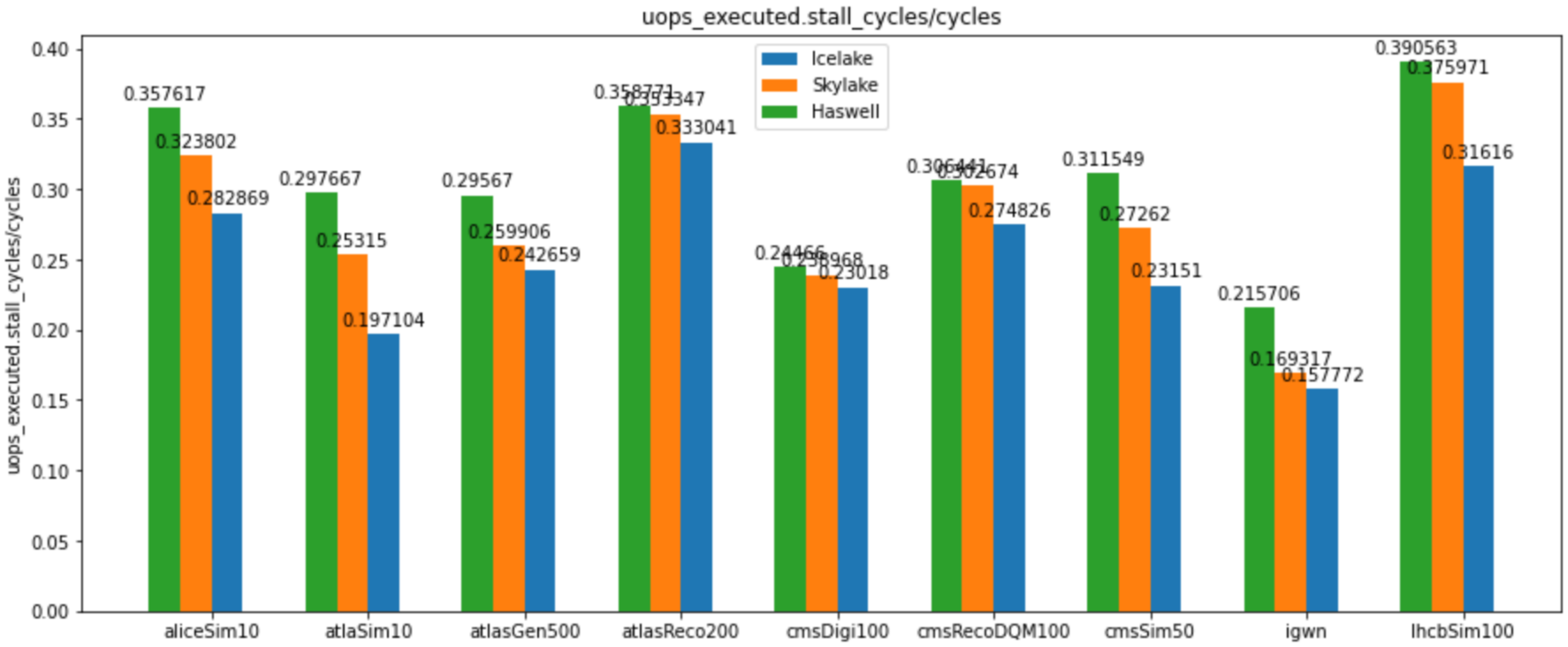
CPU efficiency IPC (max 4)

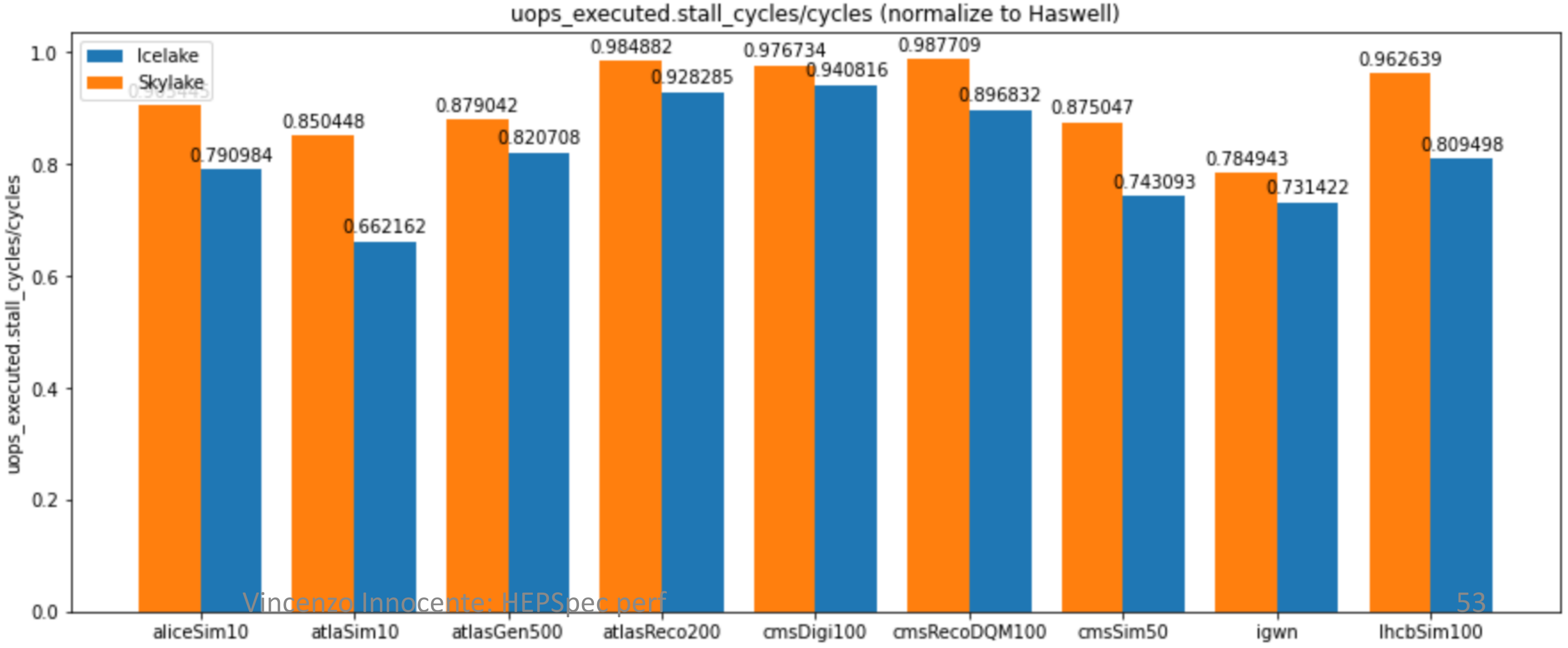




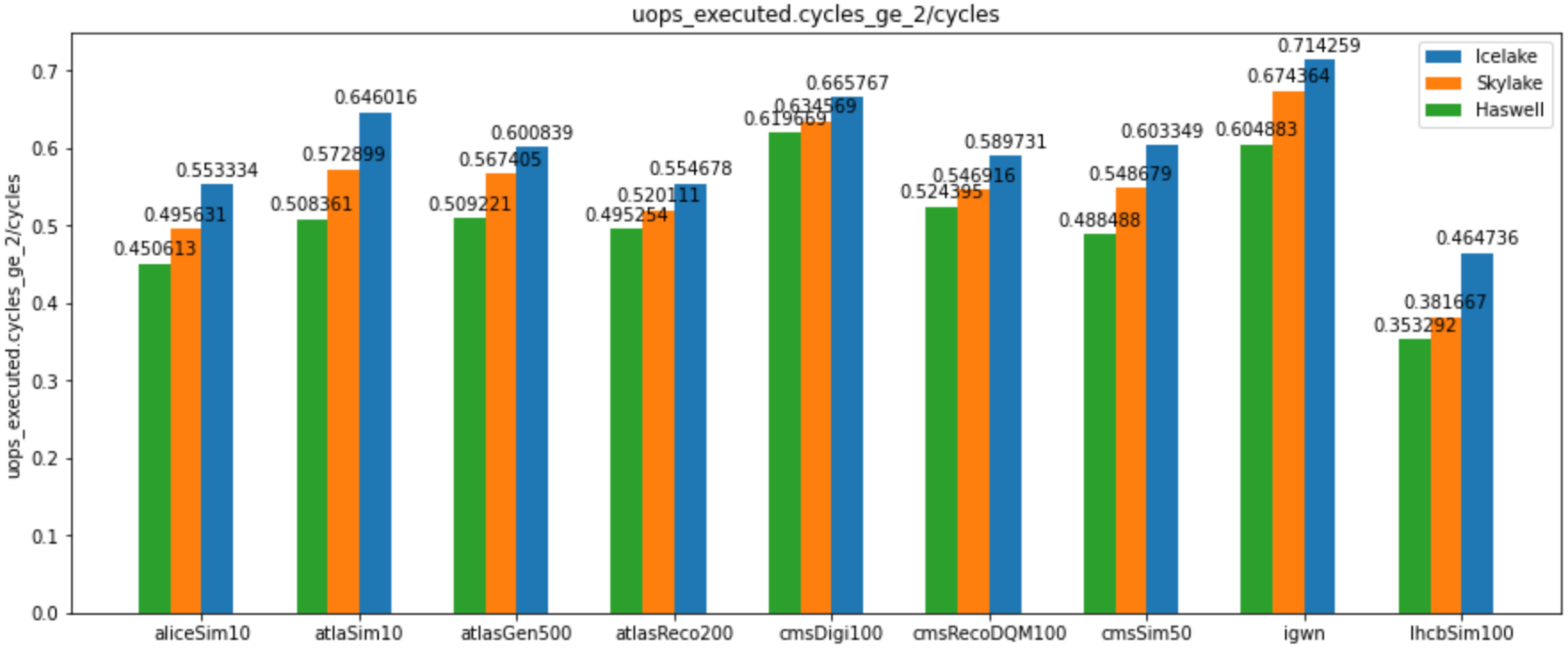
Stalls

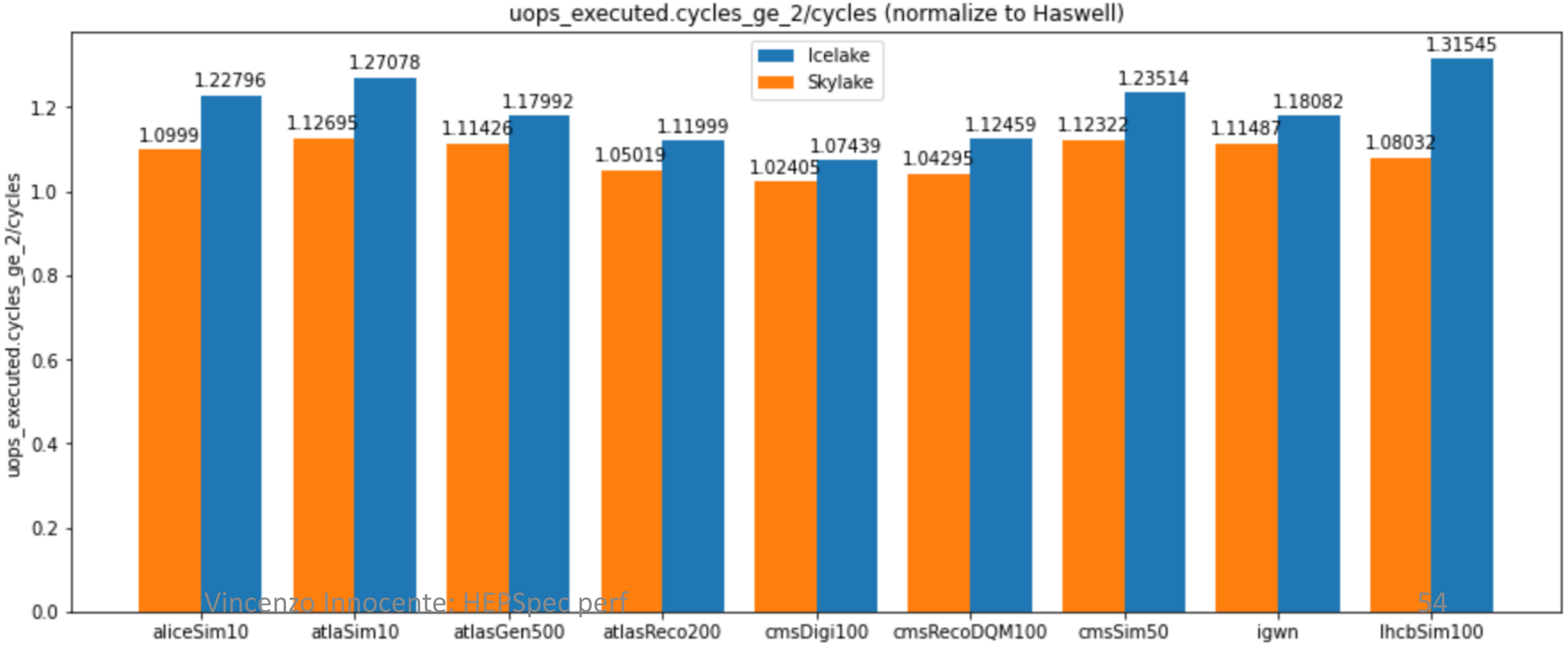
cycles where no instruction executed



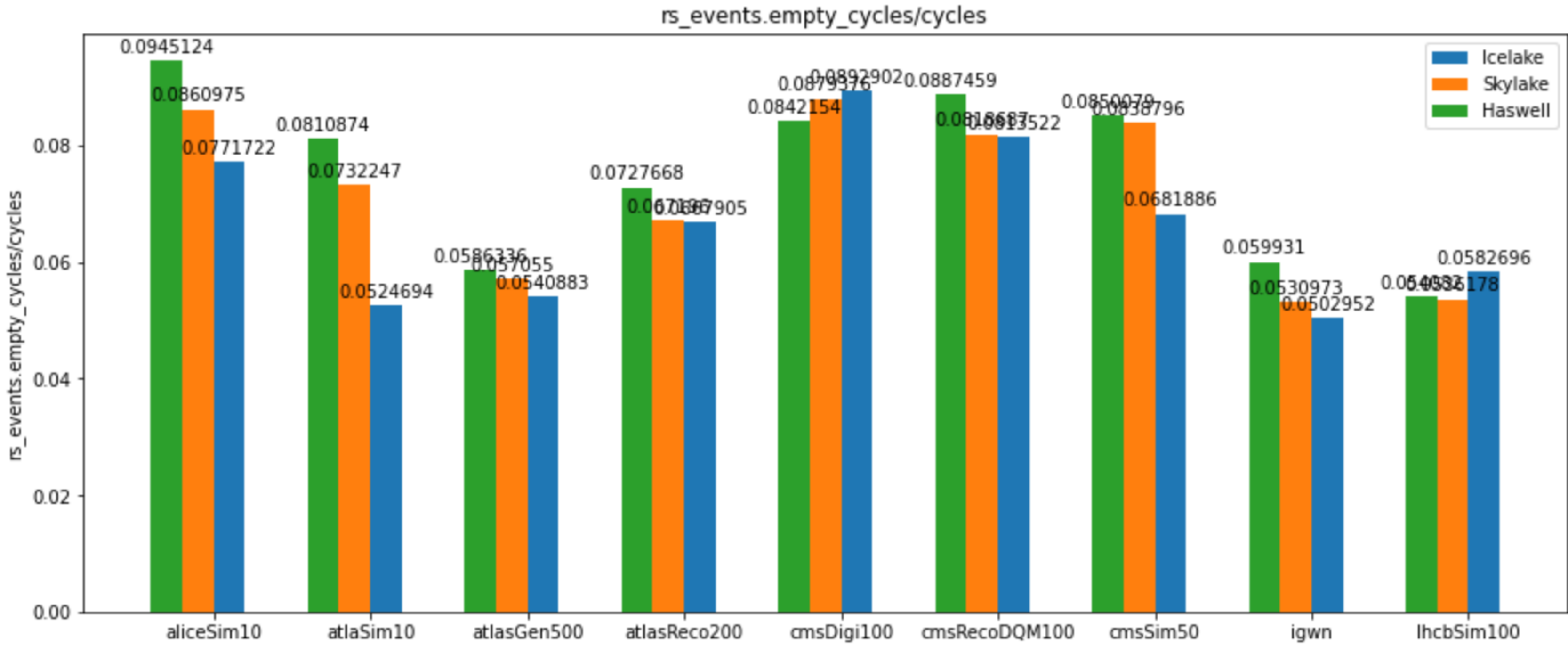


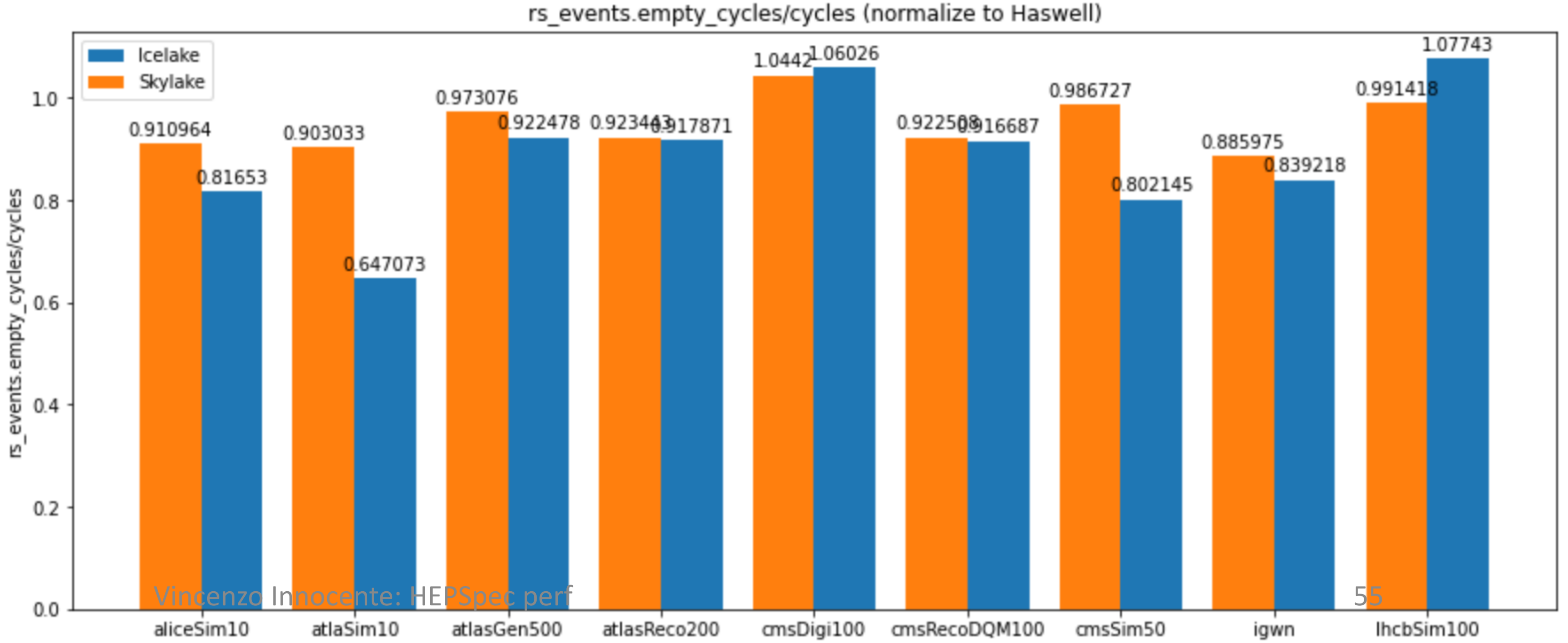
Parallelism: #cycles >2 instrs



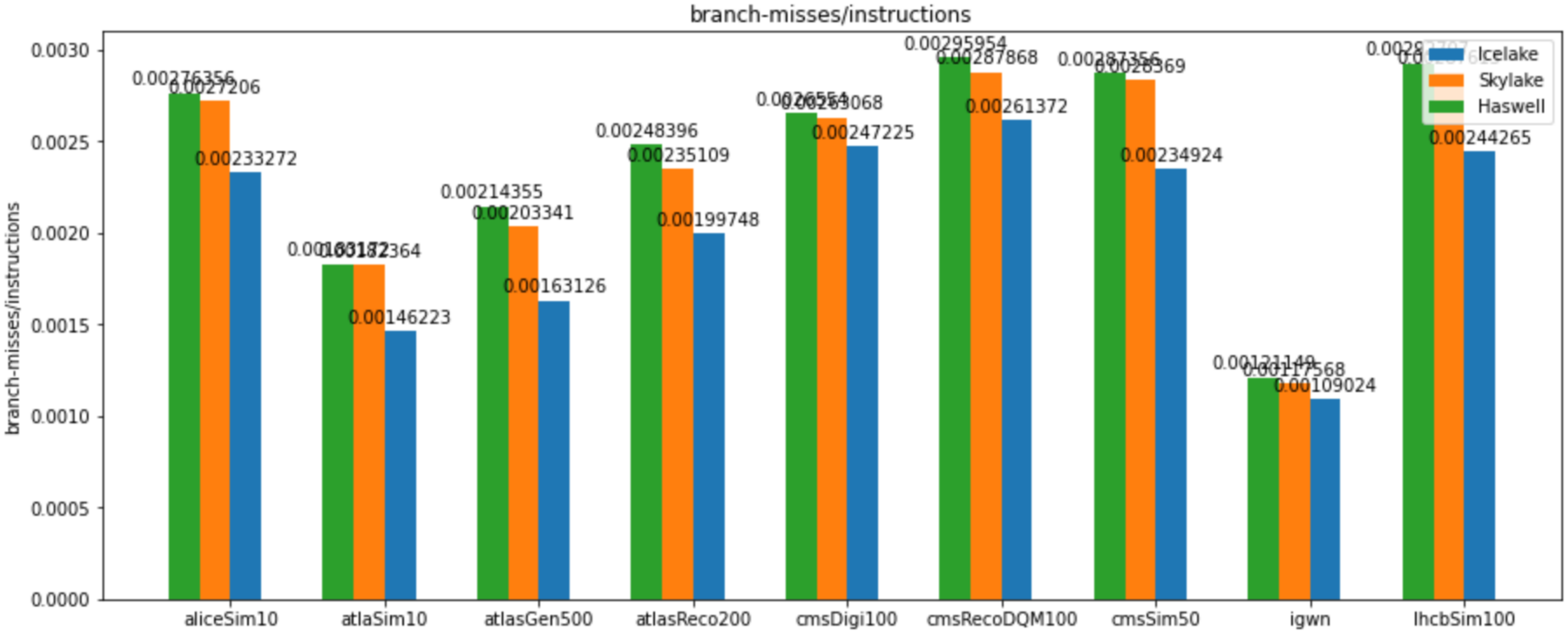


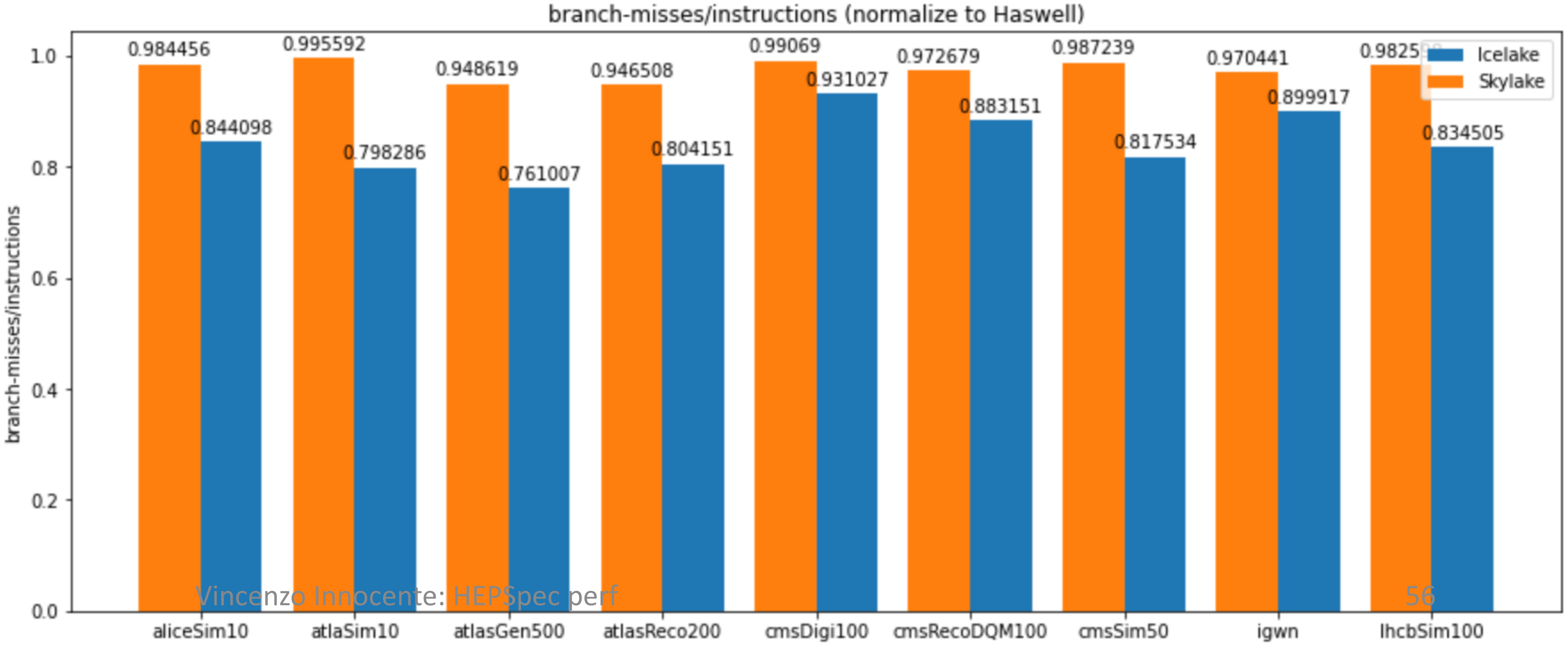
CPU starvation empty reserve-station



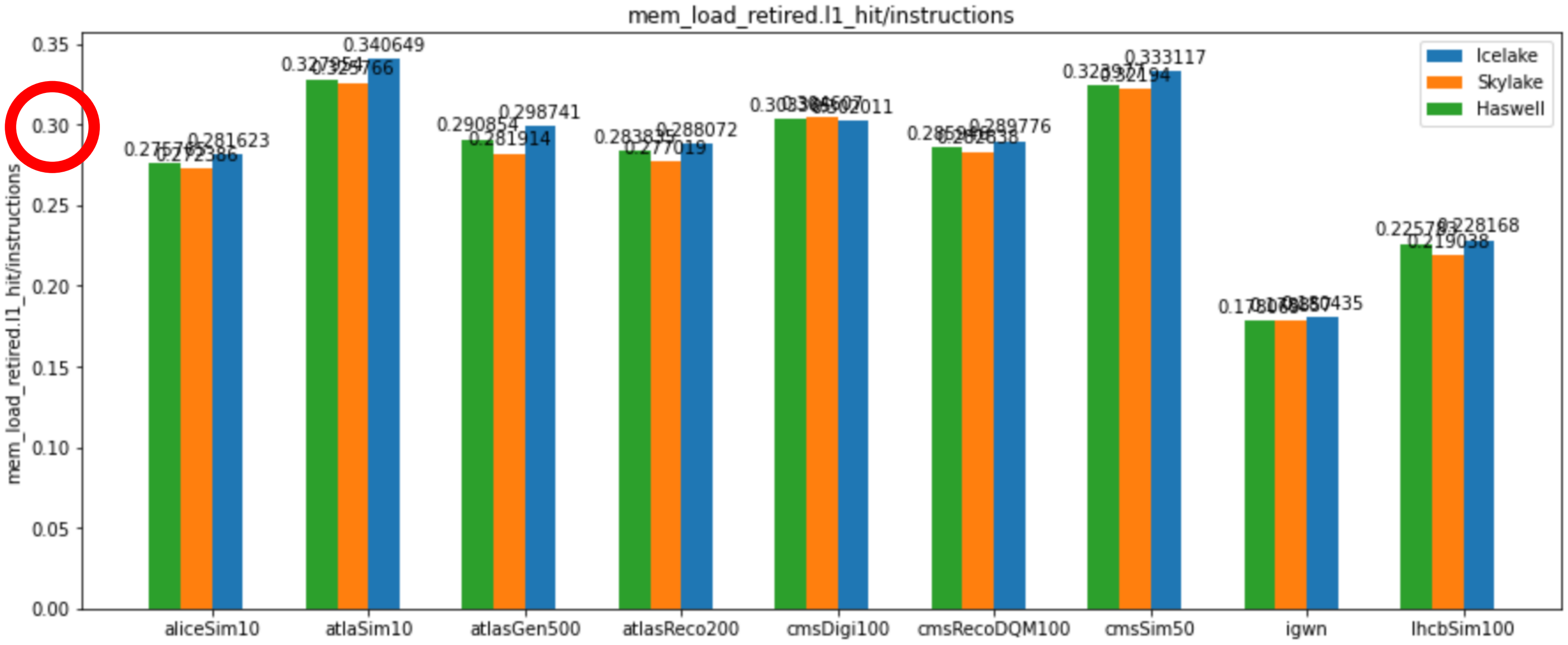


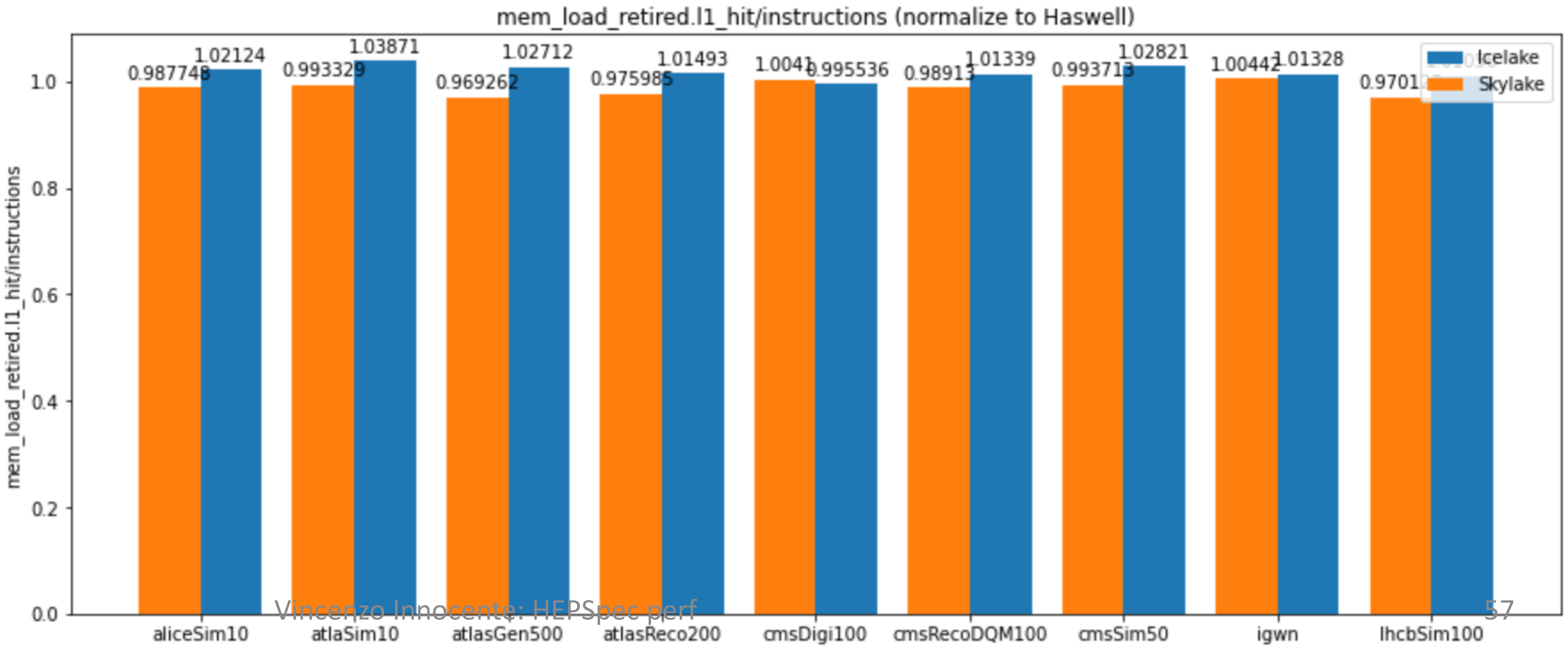
Branch misses



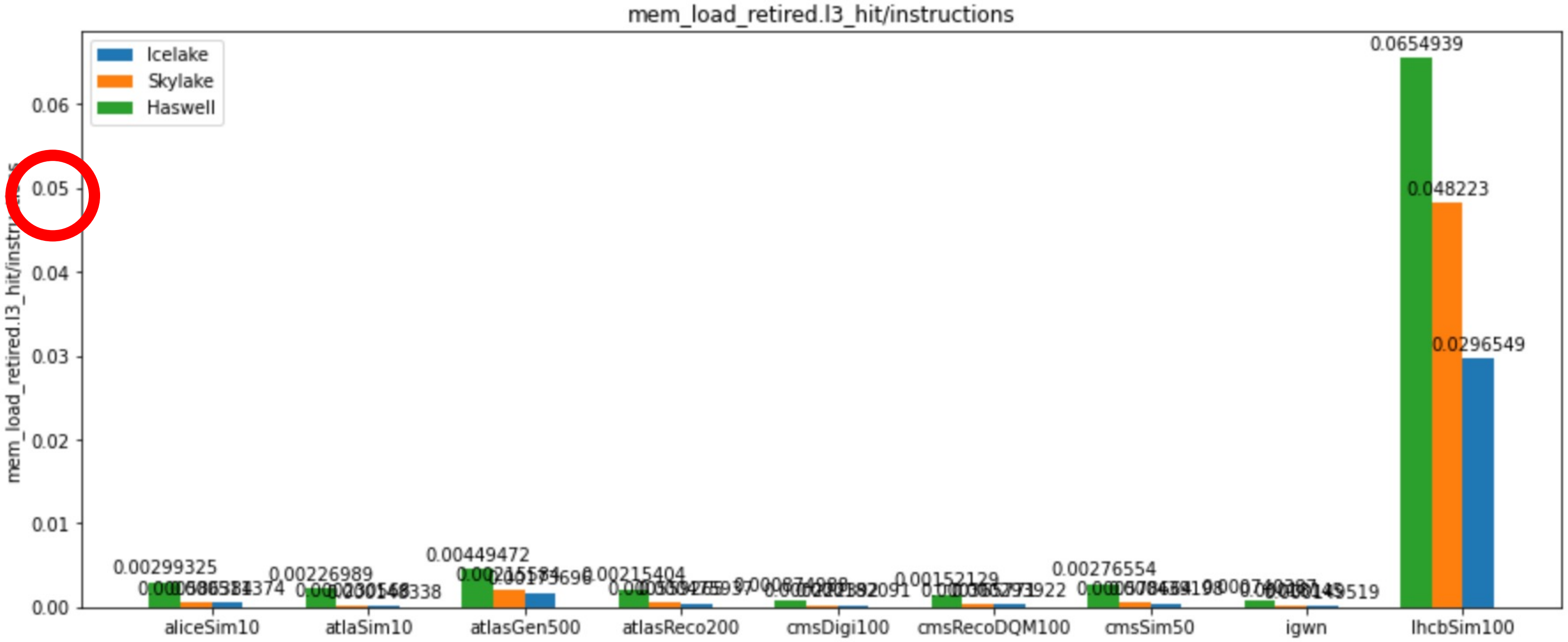


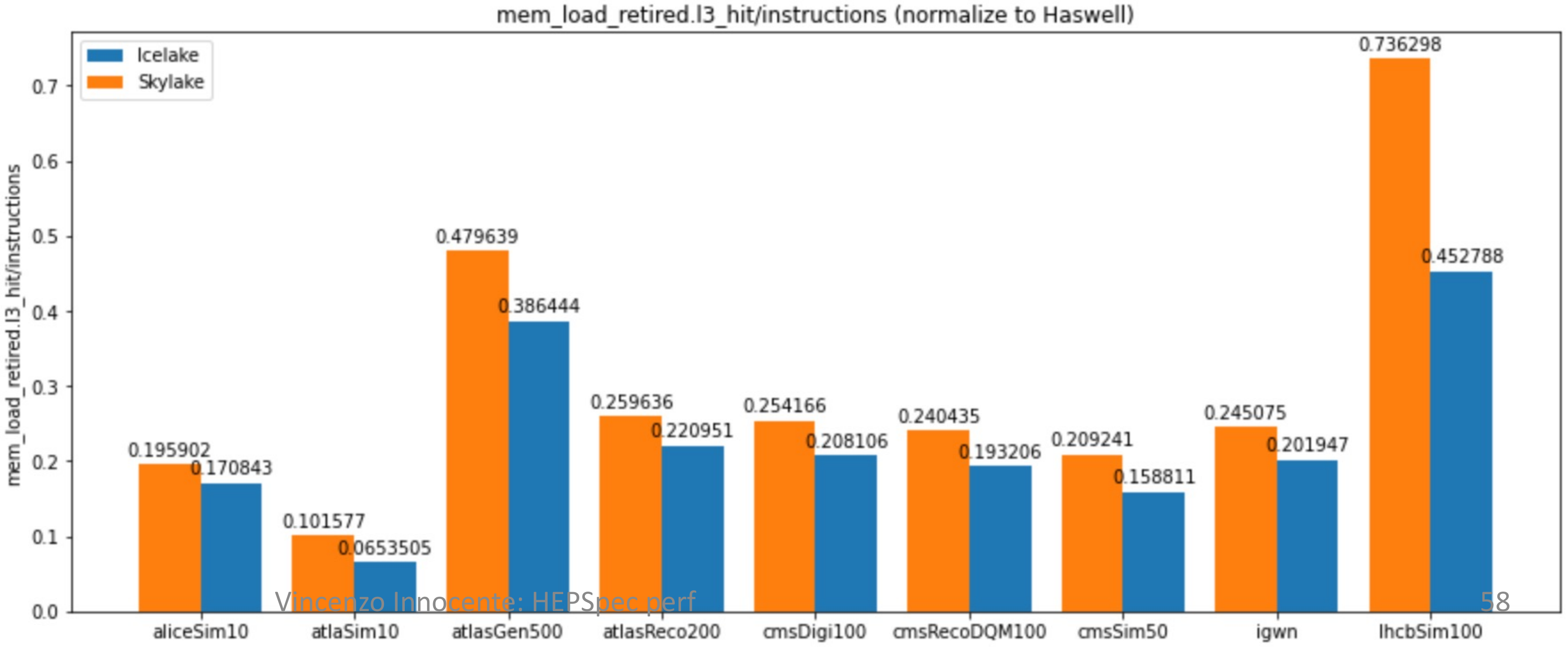
L1 cache access (4 cycles latency)

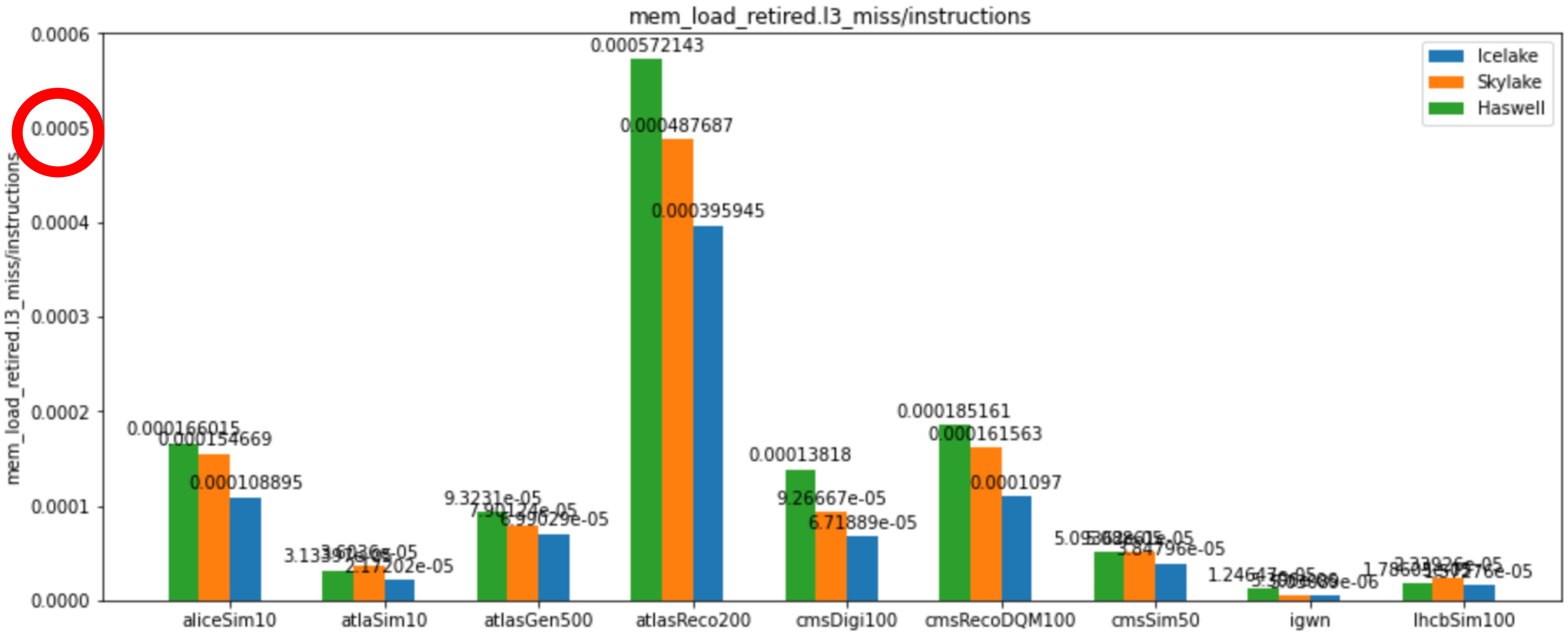




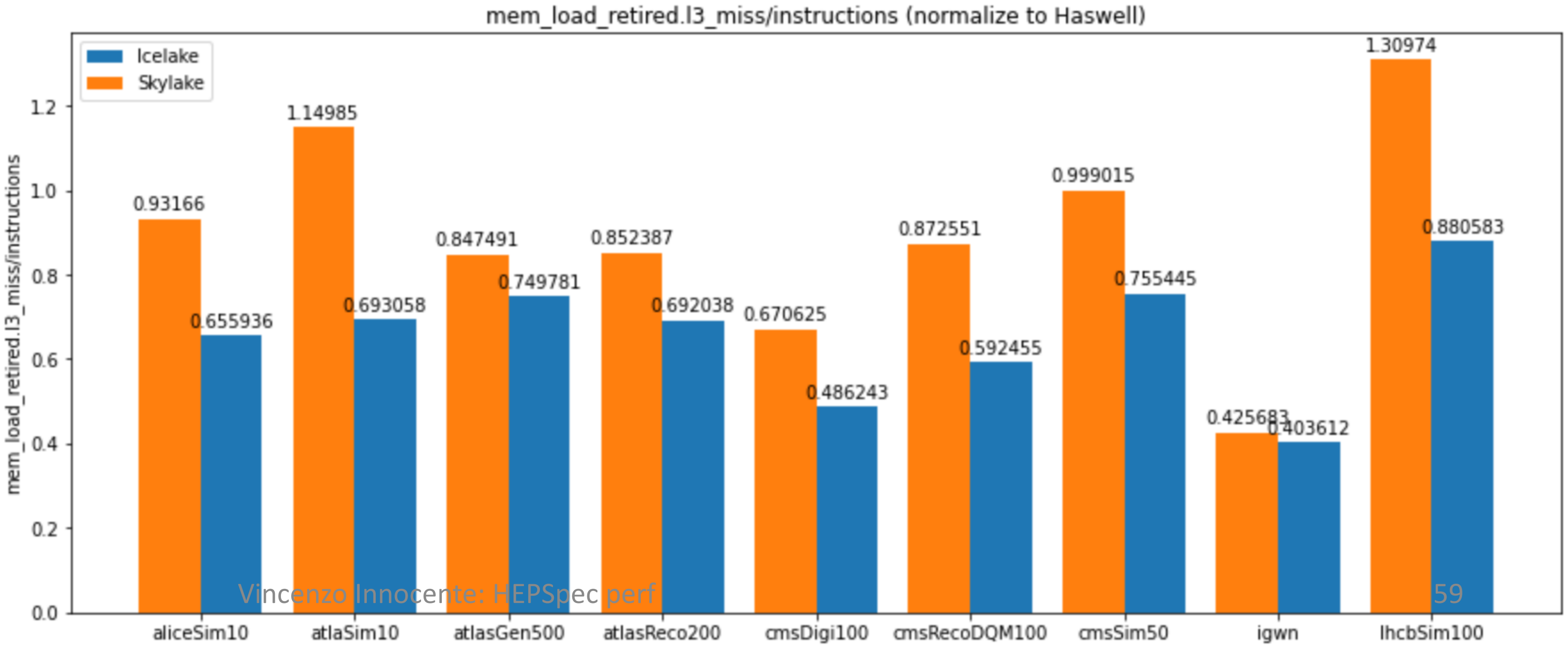
L3 cache access (50 cycles latency)



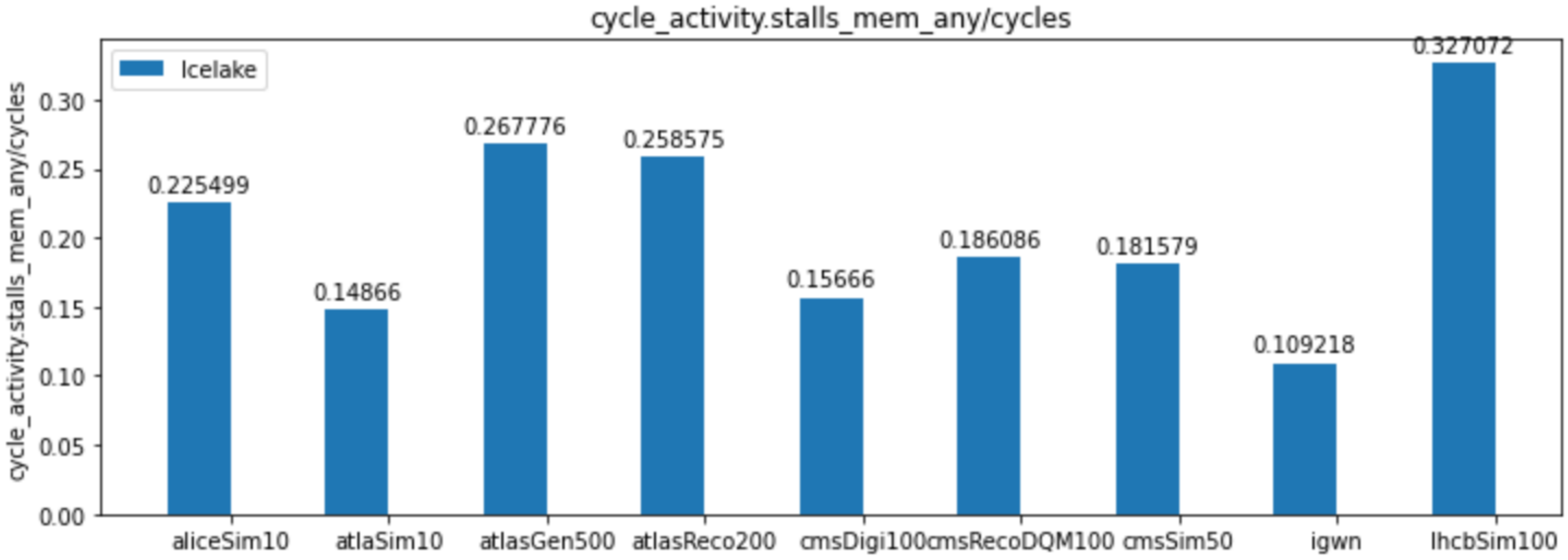




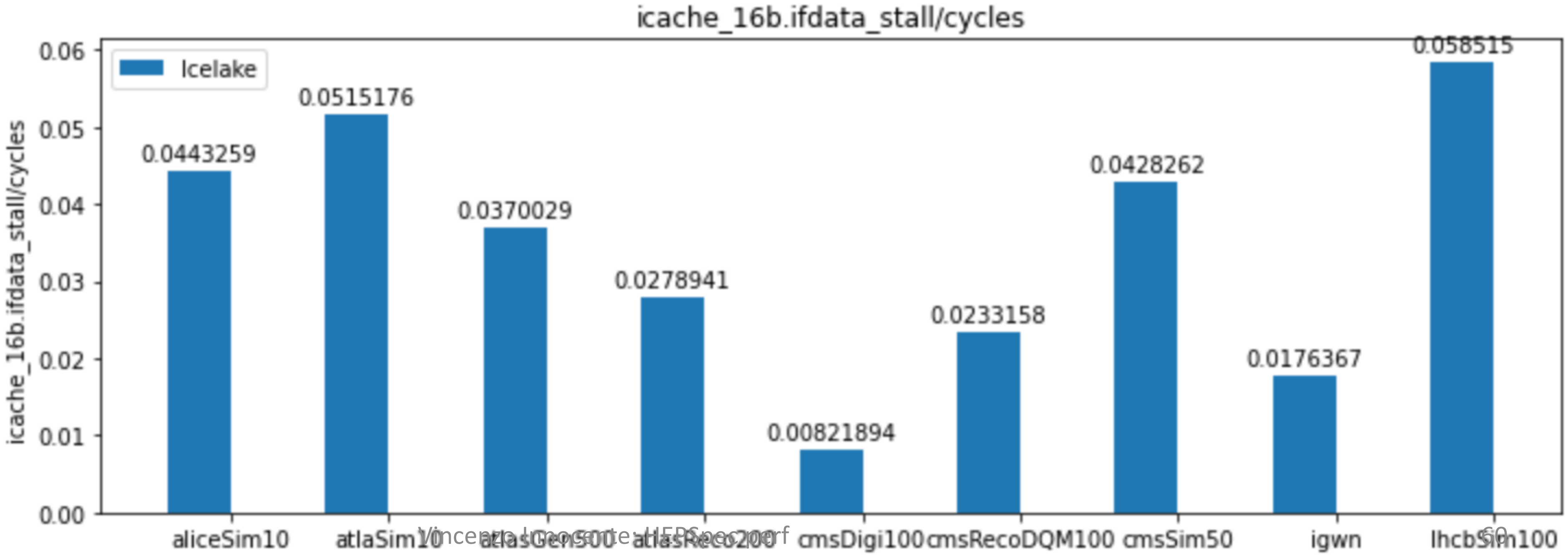
Main memory access (>200 cycles ~100 ns latency)



Stalls on memory

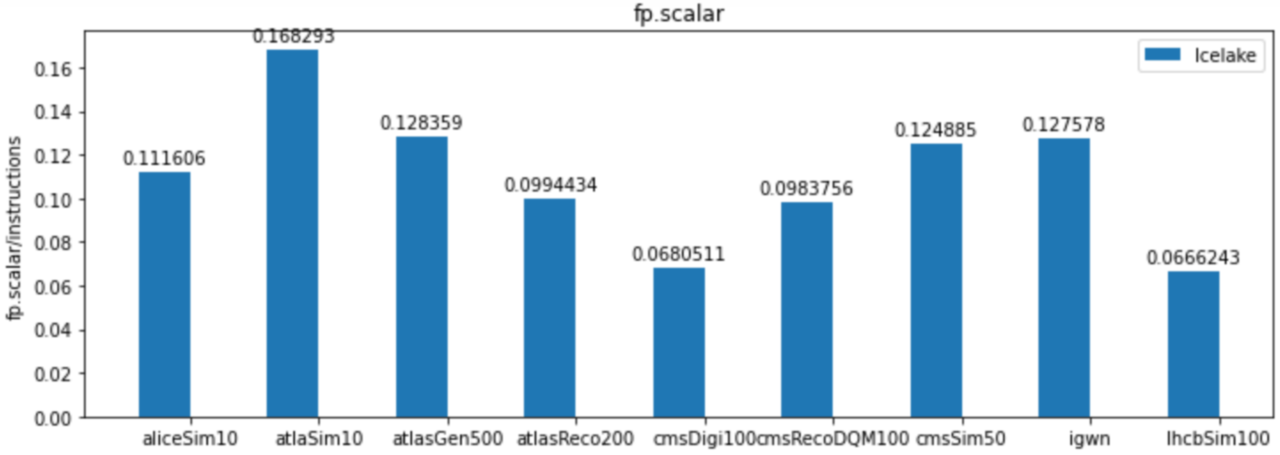


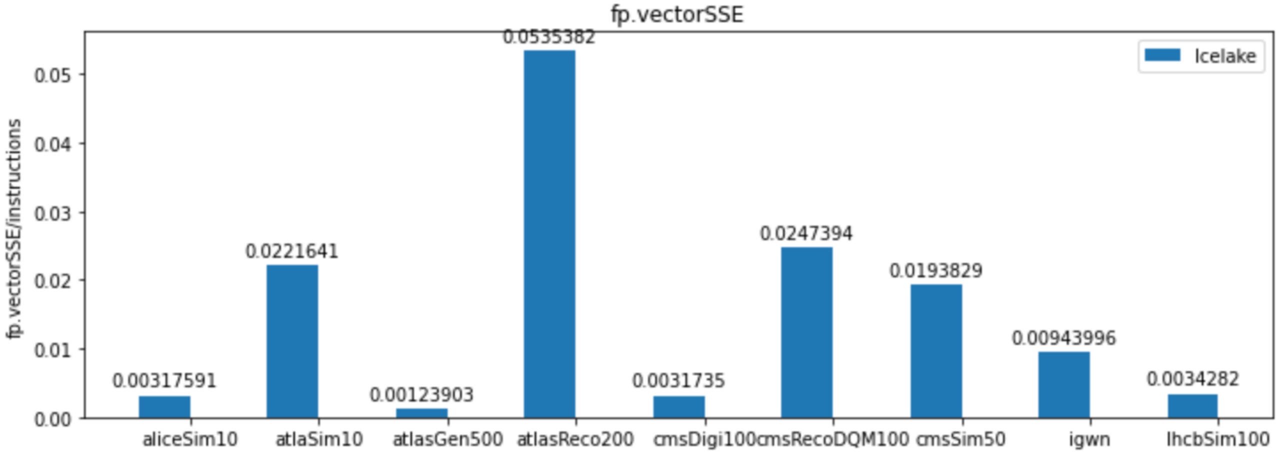
Stalls on instr-loads

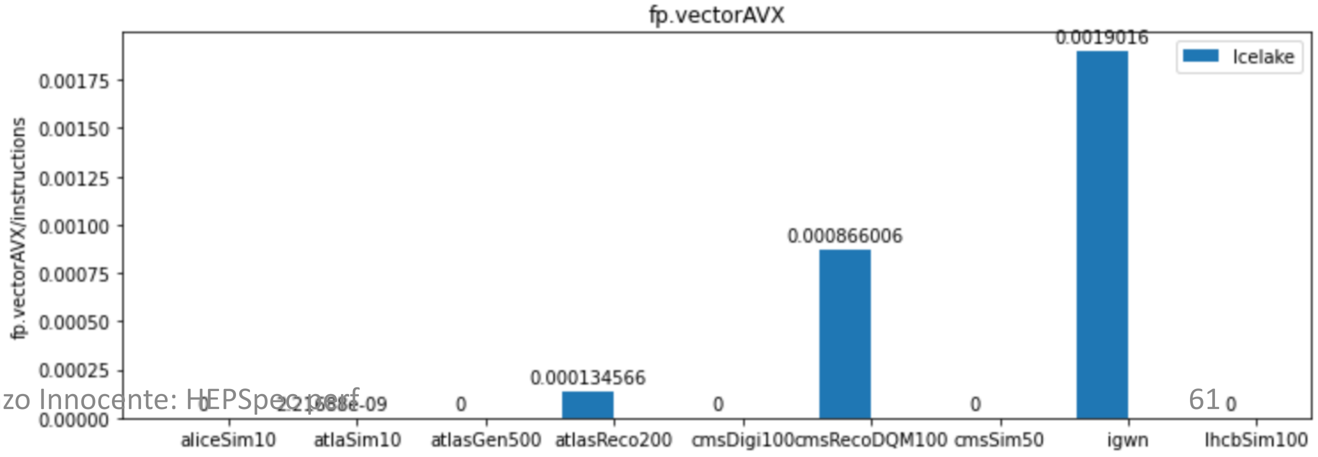


Floating-point

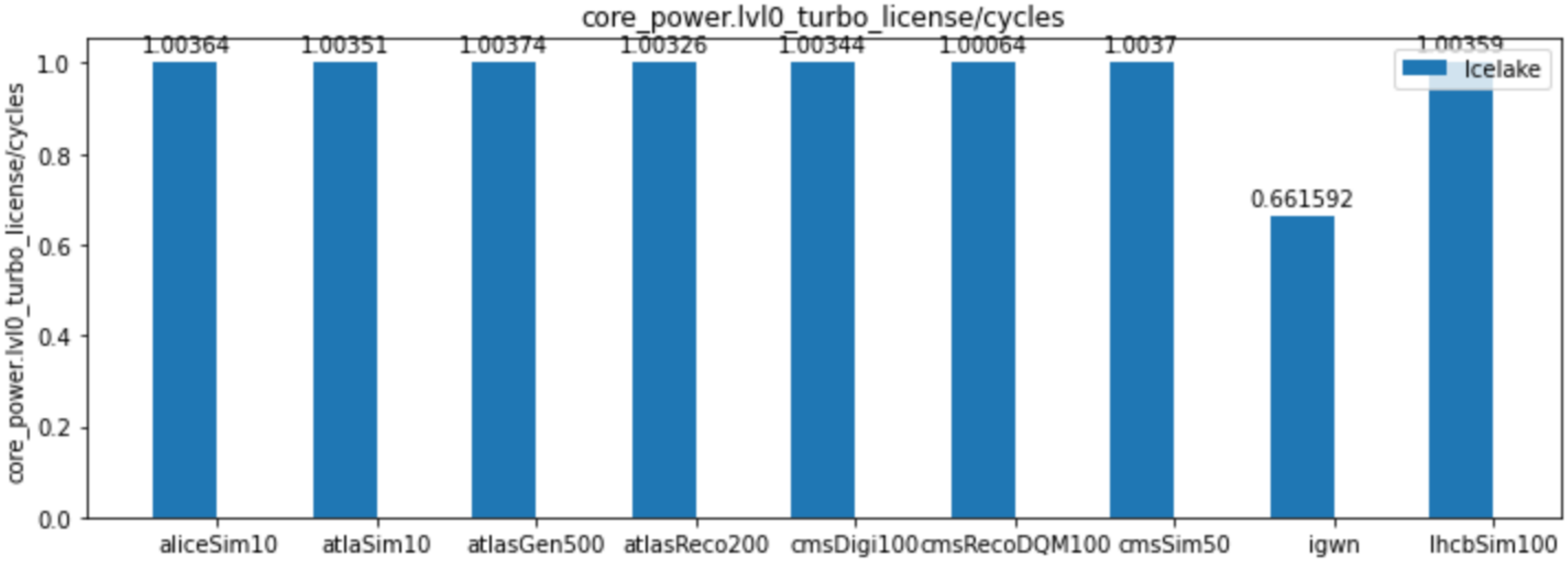
code compiled for SSE. Presence of AVX (even AVX512 for igwn) means that ”fat libriaries” are used





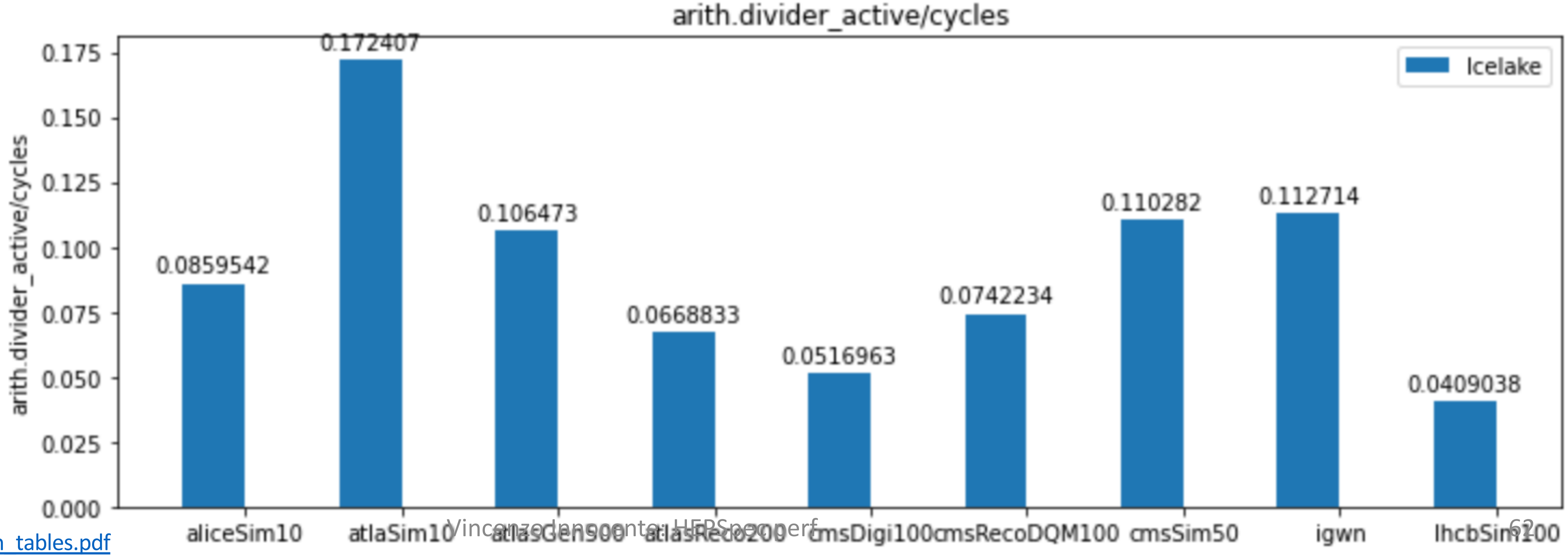


Freq throttling



divisions and sqrt

(latency: 10-20 cycles)



https://www.agner.org/optimize/instruction_tables.pdf