Research for AI Workloads & NVIDIA Nsight Users

Final Presentation

3.5.24

Contributors

Danielle Cox Phillips, Intel

Tom Rhinelander, Proximity Lab Kim Torres, Proximity Lab

Rina Doherty, Intel

Agenda

1. Project Details
2. Findings
3. Recommendations & Next Steps

Project Details

Project Goal

Deliver informative and actionable research findings of

user needs for large scale AI workload optimization tools

that enable development of enhanced versions of VTune Profiler

Top-line Story

AI model development is increasing in number and pace, but those doing it will be increasingly less technical in the traditional sense; even those with GPU expertise may expect AI platforms to manage kernel-level issues.

While model builders want to optimize workloads, they are often looking to tune it to get it “good enough” and then hand it off for deployment. They increasingly want tools that deliver smart, actionable advice, often with the issues and recommendations integrated into high-level tools.

Findings Summary

1

A desire for smart recommendations

The top desire is to get smarter recommendations on what to do when a tool uncovers performance issues (current Nsight recommendations do not offer enough value). Interviewees cited ChatGTP-like interfaces, and AI in general, as a way to uncover useful, actionable next steps.

2

Interest in core features and ease of use

Interviewees definitely see value in the timeline, memory analysis, and comparing runs. They did not care about the metric formula feature much.

The desire for intuitive, simplified interfaces, as well as integrating output (analysis and recommendations) with other tools, was noted.

3

Many future users will be less technical and about "good enough"

While not universal, there seems to be a very credible view that future users, far outnumbering those of today, will be less "in the weeds." This reinforces the need for smart recommendations and an intuitive experience.

In addition, "good enough" seems to be an evolving trend, versus squeezing every last bit of performance.

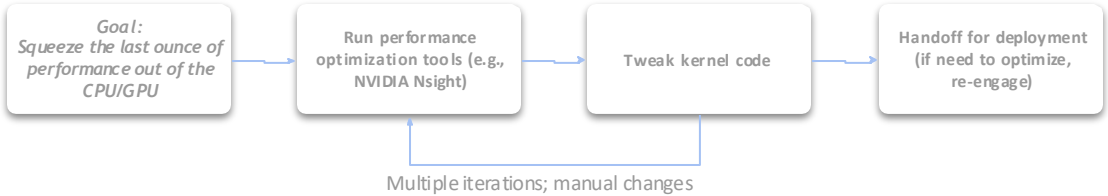
Note: These recommendations are based on 5 interviewees; we recommend both deeper one-on-one sessions (qual) with prospective users, as well as a more broad (quant) effort to validate or uncover broad trends.

The Evolving Use Case for AI Workload Optimization

Traditional user

Kernel-centric technical user (HPC, graphics, etc.); digs into the kernel to squeeze last bit of performance out

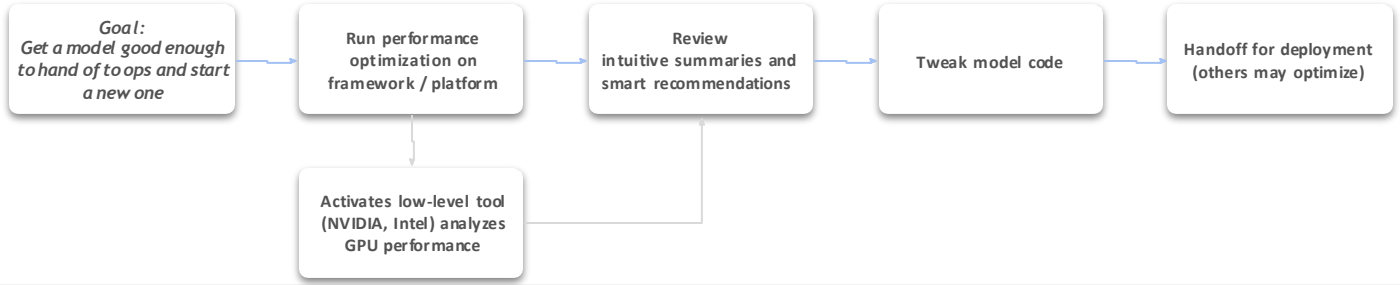
Original target persona



Model-first user

Assumes platform deals with most kernel issues; aiming for good enough; wants smart AI recommendations

What many people will likely want



AI Workload Creators are Different than Originally Assumed

Original target persona (sample titles)

- **Performance Optimization Engineer**
- AI GPU Performance Engineer
- AI GPU Profiling Specialist
- Deep Learning GPU Optimization Engineer
- GPU Accelerated AI Performance Analyst
- AI Model Inference Optimization Engineer
- GPU-Accelerated Machine Learning Engineer
- Neural Network GPU Performance Specialist
- GPU Profiler for AI Workloads
- AI Inference System Performance Engineer
- GPU Optimization for Deep Learning Models
- Application Developer

Who we found creating AI workloads

- Sr. Data Scientist
- Supply Chain Data Scientist
- Lead Software Engineer
- Data Scientist
- Machine Learning Scientist

Project Overview and Inputs

Stakeholder Interviews

Workshops with Internal Intel stakeholders helped develop an understanding of products / features, identify requirements, and clarify business goals

Details

- Serap Suvari, Product Marketing Engineer
- Vladimir Tsymbal, Development Tools Software Engineering Manager
- Julia Fedorova, Principal Engineer, Software and Services Group
- Nishant Agrawal, Senior Architect

Qualitative Interviews

Conversations with five people who optimize AI workloads and use NVIDIA Nsight Compute to understand their experience with NVIDIA Nsight, the user journey around AI optimization, and to uncover issues and opportunities for Intel's VTune Profiler

Details

- Qualitative Interview Participant Overview

Qualitative User Interviews Research Plan

Goal

Understand target personas experience with NVIDIA Nsight and the user journey around AI optimization, with a deep dive on experience with NVIDIA Nsight Compute (usage, perceptions, pain points, benefits, desires, etc.) to uncover issues and opportunities for Intel's VTune Profiler

Note: All interviews will be deleted from Dropbox within 6 months of interview date. Consent was obtained for all participants.

Detailed Research Plan & Discussion Guide

Research Details

- Ideal persona characteristics confirmed with Intel
- Up to **7** interviews (minimum 4, depending on recruitment)
- Interviewees recruited and paid by Proximity Lab
- **1.5** hour, recorded interviews
- Interview discussion guide reviewed with Intel prior to sessions
- Stakeholder interviews conducted to inform direction
- PL will not reveal sponsor of study
- Intel will provide privacy & consent form to sign
- *Notes: Initial recruitment was more difficult than expected, and screeners were relaxed to find participants (e.g., smaller scale of AI workloads, beyond NA geography). The initial plan was also to have a portion of interviewees screen share NVIDIA Nsight Compute use leveraging dummy data or a sandbox environment; this was found to not be possible for this project*

Persona

Technical optimization role: highly specialized development and systems engineering roles, specifically optimizing distributed AI systems

Key areas of focus

- Pain points and areas that worked well within NVIDIA Nsight experience
- Review experience of specific NVIDIA Nsight targeted features:
  - Timeline view in Nsight Systems
  - Memory analysis
  - Comparing multiple runs
  - Viewing and analyzing metric formulas

Deliverables

- Highlights & key quotes
- Recordings and summaries of all interviews
- Findings and recommendations, including gap analysis

Key Research Questions

- GENERAL
  1. **What is the typical journey for AI workload optimization?**

  2. **What are the motivations for using Nsight Compute?**
     - When is Nsight Compute used?
     - How are they using it?
     - What are the motivations?
     - What are the problems they are trying to solve?

  3. **How do people use Systems and Nsight Compute?**

- HOW TO INSPIRE USE
  4. **How do people learn to use Nsight Compute?**
     - What enables them to use the tools?
     - What support do they need?
     - What will get them up and running?
     - What would they want?
     - Where do they go for help?

IDENTIFYING VALUE

5. **What are the key Nsight Compute features for optimizing AI workloads?**
   - Most important features relative to AI workloads and scaling and how they're used ? (3-4 features)
   - How does it help them solve their problems?
   - What are the metrics they care most about?

TARGETED FEATURES

6. **What feedback do users have regarding the targeted Nsight Compute features?**
   - Timeline view in Nsight Systems
   - Memory analysis
   - Comparing multiple runs
   - Viewing and analyzing metric formulas

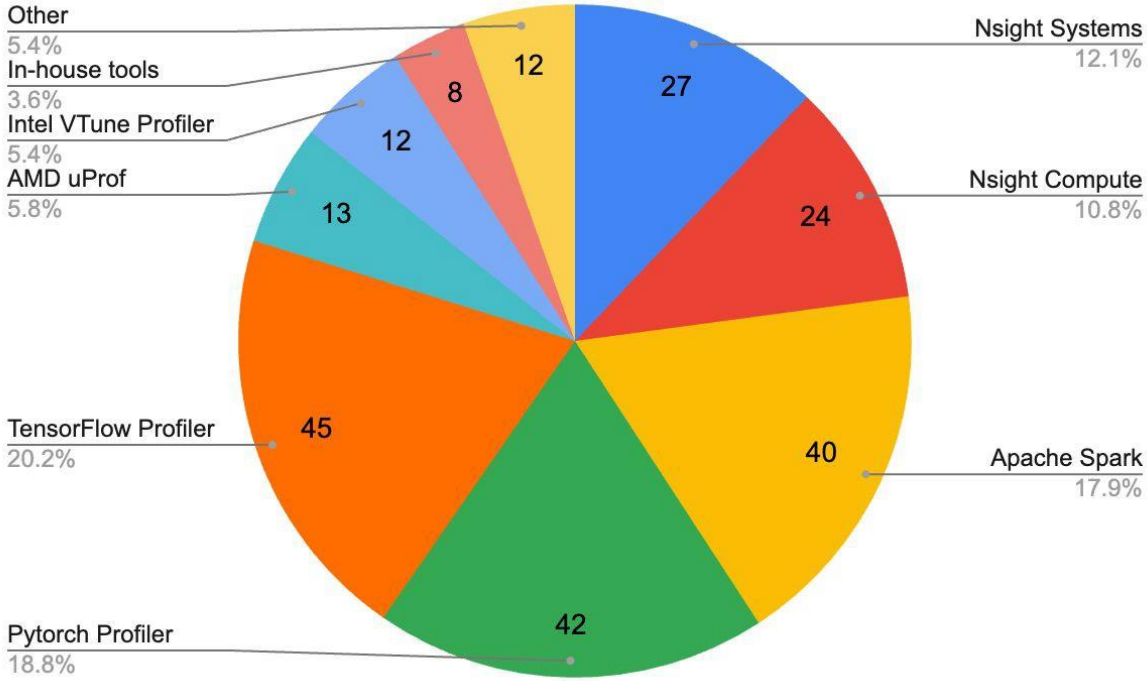
USABILITY & FUNCTIONALITY FEEDBACK

7. **What could be improved to help with AI workload optimization?**
   - What works well
   - what could be improved

Ecosystem of Respondents - General Performance Analysis Tools

Note: N=139

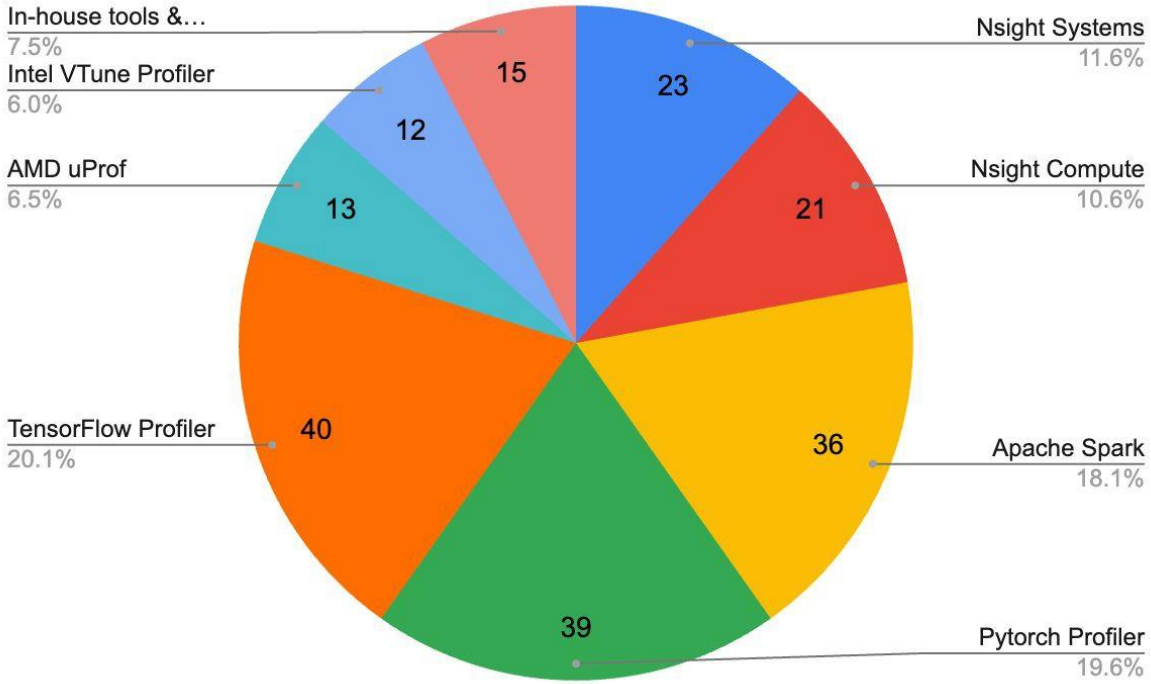
Totals represent responses collected in screener from those claiming to optimize AI workloads and were not otherwise vetted.



"Which performance analysis tools are you currently using? Please select all that apply."

Ecosystem of Respondents - Tools for AI Workload Optimization

Note: N=189
Totals represent responses collected in screener from those claiming to optimize AI workloads and were not otherwise vetted



"Which tools are you currently using to optimize AI workloads or systems? Please select all that apply."

Findings

Qualitative Interview Participants

PA

Sr. Data Scientist

Builds AI solutions to help production engineers improve overall production and reduce costs

RECORDING

PN

Supply Chain Data Scientist

Develops and deploys AI and analytical solutions at scale in conjunction with enterprise projects

RECORDING

MB

Lead Software Engineer

Builds frameworks that enable companies to use database; develops and trains models for over 5B records daily

RECORDING

CX

Data Scientist

Builds AI models for caseload forecasting and better hospital revenue; uses GPU training for transformer architecture in natural language processing

RECORDING

IN

Machine Learning Scientist

Accelerates ML model training using data parallelism and model parallelism; scales them to GPU nodes

RECORDING

Participant 1 (PA) - Senior Data Scientist

Zoom session on 2/7/2024

Recording

Note: All interviews will be deleted from Dropbox within 6 months of interview date

Role & Background:

- Building's tools to help engineers improve production and reduce costs
- Team of 4 data scientists focused on performance analysis
- Works closely with solutions architect team, infrastructure team, and Sec Ops; meets with business owners

Tools used to optimize AI workload:

- General performance: Apache Spark, Nsight Compute, TensorFlow Profiler
- AI workloads: Nsight Compute, TensorFlow Profiler, cuDNN

Use Case & Scaling:

- Running simulations every day; 3 years of data (massive amount); 8-10 nodes
- On-premises optimization is most important: 80% of AI workloads in cloud and 20% on-premises; trend to keep more on-premises for security
- Looking to squeeze as much out of GPUs
- Optimize CUDA kernel to ensure stability

NVIDIA Nsight Learning & Training:

- Doesn't use support (takes multiple days for response and sales is involved)
- Learned to use from YouTube tutorials from people who designed Nsight Compute
- NVIDIA support not helpful since they are junior developers
- Steep learning curve to understand Nsight Compute

NVIDIA Nsight Use:

- Mostly using Nsight Compute; only quarterly use of Systems
- Using Nsight Compute for past 3 years, 2-3 times per week
- Use Nsight Compute when identifying a bug or longer run time, bottlenecks
- Copy/paste section of CUDA kernels and compare 2 to help answer questions
- Perception that Nsight Compute helps to get work done in less time

- NOT HELPFUL
  - Timeline view: Too high level; doesn't communicate any value
  - Don't care about viewing and analyzing metric formulas
- PAIN POINTS
  - Would like it to be more customizable (access to raw data)
  - Need more timely predictions
- WISH LIST
  - Want fewer charts and more granular reporting
  - AI recommendations based on frequent use / combine views
  - Smart chat to answer questions in plain english
  - Sandbox environment to learn
  - Customizable alerts
  - Comparison of code changes for iterations to improve and what were made in same view

Compute due to better features like comparison, granular details, customizable, actionable metrics, memory analysis, speed of light metric changes

Participant 2 (PN) - Supply Chain Data Scientist

Zoom session on 2/7/2024

Recording

Note: All interviews will be deleted from Dropbox within 6 months of interview date

Role & Background

- Collaborate with, but only discuss budget and context with business users
- Optimization increasingly important as neural networks increase
- Nsight is their exclusive tool for GPUs (using NVIDIA GPUs)

Tools used for performance analysis and optimization of AI workloads: Nsight Systems & Compute, Apache Spark, Pytorch Profiler, TensorFlow Profiler

Use Cases

- Not just interested in squeezing the GPU, but also in determining when a model is good enough and how to save time for people
- Non-traditional, more tabular in nature(convert tabular data into images)
- Do most optimization on-premises for initial dev and then move to cloud
- Easier to troubleshoot on-premises; much simpler process than cloud
- Need to optimize is increasing; need to reimagine flow every 3 months
- Pace of new features is overwhelming and they are hidden

NVIDIA Nsight Learning & Training

- Need to really understand the language before using or user will be lost
- Best resource was “Fast AI” and YouTube tutorials for learning; more tribal knowledge sharing

NVIDIA Nsight Use

- Use until it's ROI starts to diminish; don’t want to spend a lot of time in Nsight
- Use Systems more frequently than Nsight Compute since it is higher level to catch issues, then dig into Nsight Compute as necessary
- Mostly comparing iterations; make changes one at a time and then re-test; 10 or more iterations
- Use is increasing as pressure to save money increases

NVIDIA Nsight Features

- VALUE: Appreciate identifying repetitive cycles in the timeline view; ability to hover and see issues; speed of light metric, memory analysis, comparing runs
- NOT HELPFUL
  - Nsight Compute and Systems don’t look and feel the same
  - Not interested in viewing or analyzing metric formulas
- PAIN POINTS
  - Need to identify the root cause, not just symptoms of issue ; user needs to spend a lot of time figuring out what to do next
  - Too manual
- WISH LIST
  - Generative AI; search chat bot; would like it to search for issues, offer suggestions, and then accept
  - AI recommendations based on use and analysis; nudge in the experiments in background
  - More user friendly so others can get info without learning
  - Would like guidance about best practices

Participant 3 (MB) - Lead Software Engineer

Zoom session on 2/8/2024

Recording

Note: All interviews will be deleted from Dropbox within 6 months of interview date

Role & Background:
- Builds Tframeworks and platform that enables companies to use database
- Training over 12 mo. delsdaiy over an 80-node data processing cluster

Tools used to optimize AI workload:
- General performance: Apache Spark, Nsight Compute, TensorFlow Profiler
- AI Workloads: Apache Spark, Nsight Compute, Pytorch Profiler, TensorFlow Profiler

Use Cases:
- Allcloud-based (AWS and Google Cloud); varied hardware infrastructure; some Intel, some NVIDIA
- Want to be able to train different models, predict future performance, and scale
- Use AI to predict behavior of database and give insight about building queries
- AI is taking place of database expert to give insight on structuring queries

NVIDIA Insight Learning & Training:
- Orghas a Slack integration for NVIDIA support, which they don't use often, but is helpful
- Learned product through meetups and tech seminars

NVIDIA Insight Use:
- Start with reviewing Systems and then dig into Nsight Compute to get to deep kernel level analysis
- Use Nsight Compute to identify how model is performing; focus on how to make the framework better for scale
- Use Nsight Systems and Compute at the beginning of modeling to understand how code will perform and iterate
- Use both Nsight Systems and Compute on a weekly basis, always before a new iteration of a model

NVIDIA Insight Features:
- VALUE: detailed analysis with reporting and summary with timeline, accessing lines of code, memory analysis
- PAIN POINTS
  - Where data is being cached; information about memory blocks
- WISH LIST
  - Have an iterative view to understand what has been changed on each version
  - Would like recommendations for what to change on each iteration to optimize performance
  - AI recommendations for flagging issues and identifying what actions should be taken
  - More summarized view to help understand underlying information

Participant 4 (CX) - Data Scientist

Zoom session on 2/12/2024
Recording
Note: All interviews will be deleted from Dropbox within 6 months of interview date

Role & Background
- GPU training for transformer architectures in NLP
- Focus on model optimization, engineer focused on systems optimization; collaborate with engineer during optimization, overlap between the two

Tools used to optimize
- Gen Performance: Apache Spark, Nsight Compute, PyTorch Profiler, TensorFlow Profiler
- AI workloads: Nsight Systems, PyTorch Profiler, TensorFlow Profiler

Use Cases
- Typically running 10 experiments in parallel; optimize before deploying to smaller server
- Hybrid of cloud & on-premises; helpful for scaling up; keep PHI on-premises to minimize risk

NVIDIA Nsight Learning & Training
- Team knowledge sharing is key way of learning; learning by doing
- Occasional reference of NVIDIA developer documentation or forums, Stack Overflow

NVIDIA Nsight Use
- Uses both Nsight Systems and Compute, but Compute more often (1-2 times per week)
- Systems will be 2nd or 3rd step in flow for diagnosing issues
- Run profiler in background until job completes and then analyze
- Nsight Compute used from POC to monitoring system traffic
- If getting generic error or intermittent error, use Nsight Compute to troubleshoot

NVIDIA Nsight Features
- Likes GPU Trace and Profiler, visualization and reporting of events , troubleshooting performance
- PAIN POINTS
  - Takes time to learn new features
  - Set up of GPU Trace and Profiler is difficult
  - Takes a long time to troubleshoot, error messages are focused on protocols, but do not provide direction
- WISH LIST
  - Would like sandbox environment for learning
  - Simplify the experience for ML users; help to slice data
  - Better education for users to understand how to use and find value
  - Would like to set timers and set different durations for different events (microsecond might not be right unit of time)
  - Guidance for next steps to resolve issues, but need to be accurate
  - Just want it to run in the background
  - Prioritize what is most important in view for user
  - Would like to be able to combine views

Participant 5 (TN) - Machine Learning Scientist

Zoom session on 2/12/2024
Recording
Note: All interviews will be deleted from Dropbox within 6 months of interview date

Tools used to optimize:
- Nsight Systems, Nsight Compute, Visual Studio (addon that can provide Nsight statistics), PyTorch Profiler, Weights and Biases Profiler (wrapper)
- Model training exclusively GPU based; all NVIDIA, but moving to other GPUs
- Collaborate with researchers in domain to make sure models are useful
- If spreading data across multiple GPUs, need to try to understand which parts are most memory and compute intensive
- YouTube tutorials from NVIDIA developers are most helpful
- NVIDIA forum is a secondary reference

NVIDIA Nsight Learning & Training:
- Hands on demos / learning by doing

Use Cases:
- Mostly on-premises; many datasets have privacy concerns

NVIDIA Nsight Use:
- Primarily using Nsight for non-AI applications (CUDA kernels for other scientific applications, mapping problem)
- If using NVIDIA GPUs, don't think you need to look at CUDA kernels because frameworks have already been optimized; only when adding more parameters
- Looking for bottleneck
- PAIN POINTS
  - Overwhelmed with options and information
  - The way runs are grouped is confusing
- WISH LIST
  - Would like suggestions, more informed summary
  - Integrate AI models with profiler for more seamless experience
  - Would like better visualization of multiple runs

NVIDIA Nsight Features:
- VALUE: Memory optimization

Role & Background:
- Develop small and large scale AI models
- Some models start from scratch, others copy code from other applications

Key Research Question

What is the **typical journey** for AI workload optimization?

AI Workload Optimization

Hybrid tools are most common; on-premises solutions are critical for privacy and security

- Many data sets have privacy concerns and cloud is expensive, which leads to increase on-premises
- AI workload optimization is often done initially on-premises and then moved to cloud; troubleshooting is cited as easier on-premises for initial dev before moving to cloud

Model training is almost exclusively GPU-based

- Other apps are CPU based, but not AI models
- CUDA kernel optimization more of a concern when you develop from scratch, which isn't often the case with AI models

Frequent collaboration with other teams, including engineers and business users

- Focused primarily on model optimization, while engineers focus on systems engineering, but overlap between the two
- Collaborate with researchers, business users, solutions architects, SecOps to ensure models are optimized and providing value

After initial optimization, only reviewing performance when there are changes

- Some models start from scratch, others copy code from other applications, which requires optimization testing
- Periodically need to reimagine flow, which requires additional optimization and analysis

AI Workload Optimization

Hybrid tools are most common; on-premises solutions are critical for privacy and security

- Many data sets have privacy concerns and cloud is expensive, which leads to increase on-premises
- AI workload optimization is often done initially on-premises and then moved to cloud; troubleshooting is cited as easier on-premises for initial dev before moving to cloud

Model training is almost exclusively GPU-based

- Other apps are CPU based, but not AI models
- CUDA kernel optimization more of a concern when you develop from scratch, which isn't often the case with AI models

Frequent collaboration with other teams, including engineers and business users

- Focused primarily on model optimization, while engineers focus on systems engineering, but overlap between the two
- Collaborate with researchers, business users, solutions architects, SecOps to ensure models are optimized and providing value

After initial optimization, only reviewing performance when there are changes

- Some models start from scratch, others copy code from other applications, which requires optimization testing
- Periodically need to reimagine flow, which requires additional optimization and analysis

Key Research Question

What are the **motivations** for using Nsight Compute?

Motivations for Use

Pressure to quickly get model ready for handoff

- Perception is that Nsight Compute helps to get work done in less time; predictive maintenance is key
- Often trying to optimize before handing off to production server

Priority is to reduce training time for a model and ensure accuracy

- Most focused on model optimization, leaving systems optimization to engineers

Want to be able to train models, predict future performance, and scale

- Often retraining model challengers to replace

Interested in optimizing GPUs, but most concerned about perfecting model and saving time

- Need to try to understand which parts are most memory intensive and compute intensive when spreading data across GPUs
- Cloud solutions are not always the answer (often cost and privacy leads to on premises development and deployment)

"We use it ad hoc, but there is **more need lately with things getting more expensive.** That has turned on the pressure for us to find issues and save money."

CX
Data Scientist

Key Research Question

How do people **use** Nsight Systems and Compute?

General Use

Often using Nsight Systems for high level view, then drill down in Nsight Compute as needed

- Identifying bottlenecks through Nsight Systems and then dig into specifics through Nsight Compute, used for POC and then monitoring
- Often using other tools to diagnose issues in addition to Nsight Compute, check hardware issue, network issue etc.

Comparing runs or model iterations is most common use case; before new model iteration

- Typically comparing iterations, making changes one at a time and then retesting for multiple iterations
- Use Nsight Systems and Compute at the beginning of modeling to understand how code will perform and iterate

Review of how model is performing; focus on how to make the framework better for scale

- Sometimes running profiler in background until job completes and then analyze
- Use of Nsight Compute is helpful when getting generic error or intermittent error to troubleshoot

Some are using Nsight for other issues, but not AI applications

- NVIDIA GPUs CUDA kernels already optimized, so unnecessary unless adding parameters
- Often no specific protocols in place regarding optimization, people are trusted to use most effective tools

"There are really **2 classes of Nsight Compute users**: Software engineers, and there aren't that many GPU engineers these days; and then ML engineers."

CX
Data Scientist

"A product is as good as the demand from customers. Increasingly, these are used by those invested in LLM. **More people will have expertise like me rather than compute architecture**."

TN

Machine Learning Scientist

Key Research Question

How do people **learn** to use Nsight Compute?

Learning & Support

Learning by video and then trying

- Often watch videos for basics, but then just figure it out on their own with real world examples
- Most interested in a sandbox environment

Knowledge sharing among colleagues is the most trusted source of information

- Users typically share how they've used the tool through team meetings, meetups, and tech seminars
- NVIDIA developer video tutorials are most popular method to learn basics; NVIDIA developer documentation & forums also cited as helpful, but utilized less

NVIDIA support not helpful due to sales involved with process and slow response

- Users who mentioned NVIDIA support were often junior developers who don't really understand the product
- Don't appreciate sales being involved and would prefer to just figure it out on their own

Learning curve is steep, and users need to speak the language

- Pace of new features is overwhelming and they are hidden; requires one to be fully proficient
- Need to really understand the language before using or user will be lost

Learning - Videos & Hands On

“Watching tutorials is key, but it is so important to incorporate

learning by doing … a sandbox environment.”

CX

Data Scientist

Key Research Question

What are the **key Nsight Compute features** for optimizing AI workloads?

Most Valued Features

No single feature cited as most valuable, but enablement to troubleshoot is most appreciated

- Users most appreciate ability to dig into the weeds and obtain information, but didn't seem to love the features

Timeline view and viewing & analyzing metric formulas did not surface as key features

- Mixed responses for the timeline and no interest in viewing & analyzing Metric Formulas
- If prompted, timeline view seen as valuable (not top of mind)

Memory analysis and comparing runs were often cited as key features

- These were consistently cited as key features they appreciate, but could be improved upon

Speed of Light mentioned multiple times as a frequently referenced metric

- Often interacting with this metric to optimize workloads

Key Research Question

What feedback do users have
regarding the **targeted Nsight Compute features**?

Targeted Features Feedback

Timeline viewed as helpful in identifying issues, but missed opportunity to offer guidance

- Often used as first step of troubleshooting, especially for intermittent or performance optimization
- Scrolling, zooming and integration with events is helpful, but users would like more help understanding the underlying view

Memory analysis is critical for AI models, but difficult to parse through information

- Enables data scientist to do troubleshooting that used to be reserved for engineer
- Would like the ability to look at particular nodes and save settings for later

Comparing multiple runs needs to be better visualized

- Too much information with no recommendations or guidance; looking for AI support for recommendations

Viewing and analyzing metric formulas is not identified as a key feature

- Helpful to build trust for new users, but then unnecessary
- Would be helpful to flag any updates and would be helpful to see code changes in comparisons

"It's easy to see there's an issue, but **it takes time to figure out what the next experiment should be**. Code iteration? Reorder the loops? The user has to do a lot of figuring out."

PN

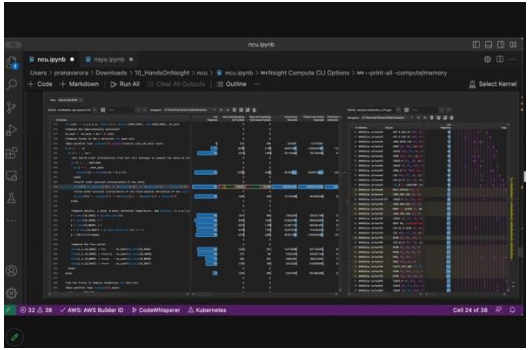
Supply Chain Data Scientist

Feedback - Comparing Multiple Runs

“This just shows output. I want to see the comparison and then see what caused the change. I can click into it, but they are separate views. **I wish they were on the same page.**”

PA

Sr Data Scientist



Key Research Question

What could be **improved** to help with AI workload optimization?

User Needs & Wishlist

AI supported recommendations based on use and analysis is the No. 1 request

- Users are spending too much time trying to troubleshoot, identify the root cause, and figure out what to do next; expect recommendations and suggestions
- Process for troubleshooting issues is too manual; would like recommendations based on frequent use and best practices

Simplification of user experience

- Users appreciate access to so much information, but are overwhelmed with options, sandbox environment to learn and plain english in the interface would help
- Steep learning curve to figure out features and capabilities; takes too much time to set up profiler and GPU Trace

Enhanced visualization to help users understand next steps

- Comparing multiple runs is confusing and could be improved for better visualization and customization
- Combining views would be helpful to decrease need to jump around

Greater customization to manage views and set alerts

- Would like to be able to access raw data, save settings; option to choose cache as well and add values
- Customizable alerts would provide better value

"It's a lot of information. For a user like me, being able to **understand what is most important** would be helpful. Help me understand what I'm looking at and what I should do."

CX
Data Scientist

"**Simplifying the user experience** is really important. I may not have the technical knowledge to go into all the details, that is why recommendations are really important."

TN

Machine Learning Scientist

Key Research Question

What **future trends**
in the field of performance
optimization should be considered?

Future Trends

As AI modeling continues to evolve, so will the expectations & use of optimization tools

- It will be critical to keep up with trends for how AI modeling is changing; understanding the key personas and their expectations and needs

Tools will need to incorporate smart recommendations to stay relevant

- Guidance is expected and smart recommendations will become table stakes
- User want to leverage what is already known, and the tool should be able to tap into that and present useful recommendations

Learning and simplification are key for new personas focused on AI workload optimization

- Optimization with AI models is critical early on, when engineers are not always managing; so it will be important to understand AI focused personas

Integration is key to deliver value where users are

- Many users do not want to spend a lot of time in optimization tools, want to have it running in the background and address as needed; would prefer to not dive into tool

- “**The way people use these tools will depend on how models are evolving.** Rather than moving to really large models, we may be gathering data on smaller models. You need to be careful about investing and be on top of the trend.”

TN
Machine Learning Specialist

- “There will be more citizen data scientists that will know less about niche topics. **Tools will have to adapt.** Computer scientists can speak the language, but **most aren’t going to care to go in the weeds.** They want to get in, make it faster and move on.”

PN

Supply Chain Data Scientist

Recommendations & Next Steps

Recommendations & Next Steps

Provide smart recommendations and better guidance for troubleshooting

- Help users identify the root cause of issues and figure out what to do next
- Offer recommendations based on frequent use and best practices
- Smart search to help users discover what they are looking for more easily

Prioritize simplification and customization

- Consider sandbox environment and plain english to help users learn
- Greater customization to manage views and set alerts
- Identify ways to surface more insights to help users not feel lost in the weeds
- Consider integrations to provide value where users are

Continued research to better understand needs for various personas

- Explore potential “citizen” user to understand needs and expectations with VTune
- Validate design concepts with potential users to ensure needs of various users are being met
- Consider UX benchmarking to identify best practices for ways to simplify user experience and provide recommendations

Feature Related Next Steps

Improvements to Timeline View

- Show CPU and GPU profiling data concurrently to provide a view of where time is spent in a given workload.
- In scenarios where CPU-GPU communication and synchronization are critical to overall performance, showing interdependencies and performance characteristics of applications.
- Refine Interactions, specifically, scrolling abilities, zooming, correlation of events, pattern identification, data aggregation and robust search, filter and sort.
- Integrate suggestions or automated insights directly within the Timeline view.
- Greater customizations for users to manage views and set alerts.

Extending Memory Analysis

- Greater customization to manage views and set alerts
- Case studies that demonstrate how to interpret the data provided by the tool and apply this knowledge to optimize applications.
- Root Cause Analysis: trace symptoms of memory usage issues back to their root causes. Provides insights into underlying reasons

Metric Formulas and Multiple Runs

- Ability to write custom metric formulas in Python.
- Show aggregate view compare runs and show what has changed from run-to-run.

Appendix

Stakeholder Interviews

STAKEHOLDER INTERVIEW

Serap Suvari — Product Marketing Engineer

Zoom session on 12/8/2023
Recording

BACKGROUND

- Over 9 years at Intel
- Working closely with Danielle and Rina

COMMENTS

- Want to focus on people doing AI; don't want all users to be HPC-centric
- Many developers don't care about deep optimization; want to focus on "low-level developers" that spend lots of time on optimization
- Need to talk to others to define "large scale, complex AI workloads"
- Use cases ("they start with two problems"):
  - Have an AI model running on GPUs
  - Already invested in GPUs, want to get the most out of it
- Successful interviews would:
  - Understand Nsight and what features they care about
  - If they use VTune, what are its strengths/weaknesses compared to Nsight
  - Understand tool usability (hear the NVIDIA is easier, so what actions do they need to take to improve VTune?)
  - Important to see what they are using (screen sharing); viewing sample code is fine

KEY TAKEAWAYS

- Seeing via screen sharing is very important (could be sample code)
- Personas
- Don't care about the full developer journey, just about performance profilers aspects
- Expanding on timeline feature:
  - Want to see CPU and GPU together?
  - See relationships over time?
  - Performance optimization engineer role
    - Focus on working on AI
    - Interested in low-level optimizations (spend a lot of time on)
  - NVIDIA Nsight Compute user is a must
  - VTune experience is a nice-to-have
  - A few could combine HPC and AI, but not most
- Want to learn
  - What functionality the care about most
  - See if that lines up with the four identified by Intel
  - If other than those four, need to learn about them

STAKEHOLDER INTERVIEW

Vladimir Tsymbal

Development Tools Software Engineering Manager

Zoom session on 12/20/2023
Recording

BACKGROUND

- Has a theoretical AI background
  - Runs AI apps on a daily basis, but only for purposes of working on Intel's tool

COMMENTS

- VTune used to be focused on HPC; AI is like HPC but with a bigger number of layers and frameworks
- AI apps bring more complexity; will need a little bit more focus on the hardware-level
- Assume analysis is happening on bigger machines (than what he uses); bigger clusters, several thousands of nodes
- VTune cannot run through analysis down to every execution on all compute nodes (thinks same for Nsight)

- Intel has a tool called Application Performance Snapshot (APS); can run on thousands of nodes on a cluster; part of Vtune (have it once you install VTune)
- It can show an overall picture; can tell how effectual it is on all the clusters; from that, the assumption it might be the same on every other node
- Also offers recommendations
- For VTune, it is important to know what are the limitations

- The metrics can help us understand how the hardware is limiting
- Common for AI applications to underutilize hardware

- VTune proving similar, but less "fancy" insight than Nsight
- Key to know where is the gap; why is it waiting? (might be communication layer, etc.)
- Might be outliers; certain parameters can give a histogram
- All nodes should have the same type of hardware (CPU and GPU)

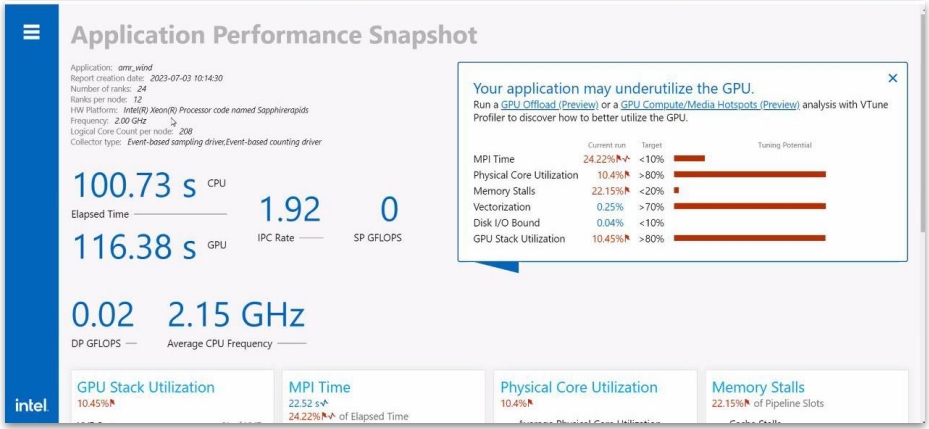
- Sometimes it requires multiple runs; Nsight can run up to 40 times to collect all sets of metrics; VTune by default using one run
- Thinks people do not want to see side by side runs; wants them in the same timeline and grid; NVidia can set baseline, and compares to that

- We believe people want to see CPU and GPU together
- Intel analyzes metric formulas; Nvidia provides the actual formula (once people work on everyday, they don't likely need them anymore)

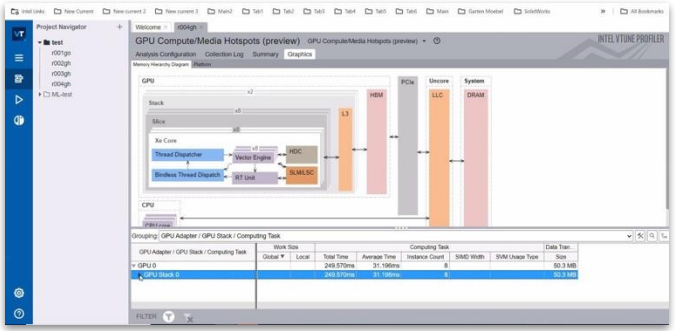
KEY TAKEAWAYS

- **Not a large-scale AI user; does not have recommendation for a dummy data load**
- Showed lots of examples of VTune/APS and talked about what users may want
- Thinks people want to see CPU and GPU together; timelines of multiple runs (not side-by-side comparisons); nice to show metric formulas (for at least FTE)
- Recommended Julia Fedorova as more of an AI expert

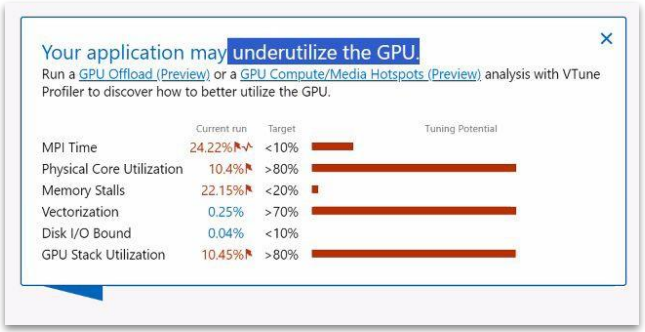
Screenshots from Vladimir Tsymbal Session 1/5/2024



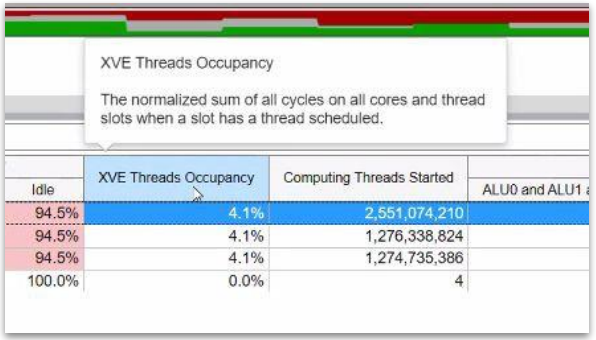
APS tool overview



Tool showing breakdown



Tool showing recommendations



Breakdown of a formula

STAKEHOLDER INTERVIEW

Julia Fedorova Principal Engineer, Software and Services Group

Zoom session on 1/5/2024
Recording

BACKGROUND
- Not an expert on AI apps, or an expert on Nsight System or Compute

COMMENTS
- Already have some understanding of what is lacking in tools; talking with Intel engineers (internal customers); very concrete what they want to see
- Don't have large scale systems internally now; but thinking of large scale, from 10s to 100s to 1000s of nodes; for definition, would use 1000 (can't "cope" with that right now)
- Would be surprised if external users (at scale) use Nsight; typically Nsight System runs on one node
- There is a tools from Meta/Facebook that we might look at

- AI workloads would likely use PyTorch or Tensor tool
- In HPC, they write custom kernels; in AI, the kernels are pre-developed and come with standard libraries

- For workloads, people typically run the tutorials with example workloads
- Be interested to see samples closer to real workloads (AI running on dozens of nodes)

- Would be interesting to see what tools they use to get an aggregated view of a workloads
- Also interesting to see what how they use Nsight Compute for AI workloads (doubts they do use it)

- Intel needs to improve its timeline capabilities
- Have seen how people use them (example was an engineer from Intel AI team)

- Very interested in how they would look at applications at scale (not clear of how they would use the tool on 100s of nodes)

- Formulas should be provided in "character view"; not in the GUI
- Don't think this is a big issue

- Comparing multiple runs is a very important feature
- Most important to know is what they are comparing (format is secondary)
- Want to learn what type of memory analysis is needed to for AI workfloads

- Also mentioned APS and how it is being adapted for AI workloads KEY TAKEAWAYS: 1000 nodes for definition of large-scale AI activity
- **Questions about users actually using Nsight for AI workloads (at all, and more than 1 node)**
- **People use tutorial and example workloads to learn (could we use those?)**

- Very interested in how users are comparing multiple runs, and wants to learn about memory analysis need

STAKEHOLDER INTERVIEW

Nishant Agrawal [redacted] age Models/Deep Learning/HPC/GPU Architecture

Zoom session on 1/24/2024
Recording

BACKGROUND
- Architectural consulting engineer
  - AI Expert, very familiar with both Nsight tools (Systems and Compute)

COMMENTS
- Prime targets would be an AI performance engineer, or solutions engineer
- Can talk to HPC people to gain some insight (not just AI)
- "Profiling" is a keyword; performance optimization is the follow up
- It's an iterative approach: Profile an app, see some gaps/places to optimize; optimize; then re-profile
- Nsight Systems and Compute have a very different way of collecting data
- Systems is designed to to understand the end-to-end run of an app, whether single or multiple GPUs; the most critical part of understanding app behavior; lightweight, compared to Compute; you get a good snapshot of what is happening
- Compute comes with a lot of overhead and takes more time for data collection; used when you want to focus on a particular section or deep dive, such as architectural analysis for a particular kernel
- There are two types of engineers (50/50 split):
  - Those focused on end-to-end
  - Another set with an app that may be very small and want it to run on a single GPU (that is where you need Compute)
- Senior engineers are supposed to do both (of the above)
- Using it once a week is a good average
- "As soon as I profile, I will take action, then redo it"
- Think they would mostly be in Systems; easy to generate data
- Timeline view is very important; want to know from users what they want to see first
  - Each CUDA stream?
  - Multiple streams in the GPU?
- NVIDIA has a lot of gaps today:
  - Comms / compute overlap
  - Memory consumption along the timeline
- Not sure how people are using the comparing multiple runs features; will be interesting to see if people use two profiles
- Analyzing metric formulas would likely be behind memory analysis and consumption overlap (in terms of user interest)
- Would like to know: If running at large scale (10 nodes with multiple GPUs), how would you understand the imbalance across GPUS?

KEY TAKEAWAYS
- The flow is profile, analyze (see issues), optimize, then re-profile
- **A split in terms of engineer use cases: Some focused on end-to-end, others focused on a single GPU**

Additional Quotes

"What's nice about Compute is it **enables troubleshooting that would traditionally be reserved for an engineer.**"

CX
Data Scientist

"With AI modeling, **I'm not reinventing the kernels,** so show me the parts you can combine and how."

TN

Machine Learning Specialist

"The **timeline view is most useful in Systems.** It's like Google Maps. It shows us where we are, and then we need to figure out where to go from there. "

PN

Supply Chain Data Scientist

"Most of the comparisons we do are baseline and iteration. We see improvement, and we go with it. I want to **pick and choose runs to view, not see them all** at the same time."

PN

Supply Chain Data Scientist

"NVIDIA has a reputation for being esoteric. For people dedicated to machine learning, it would really help to **simplify the experience**."

CX

Data Scientist

"It would great to **have some direction** about how to fix the issue, and it **must be trustworthy**. If it gives guidance and it is actually a different issue, that really hurts trust."

CX
Data Scientist

“I wish I didn’t have to worry about Compute. I miss the old CPU days, when I didn’t have to worry. Now, we’re in a different world. **It’s an art and there’s a steep learning curve**. ”

PN

Supply Chain Data Scientist

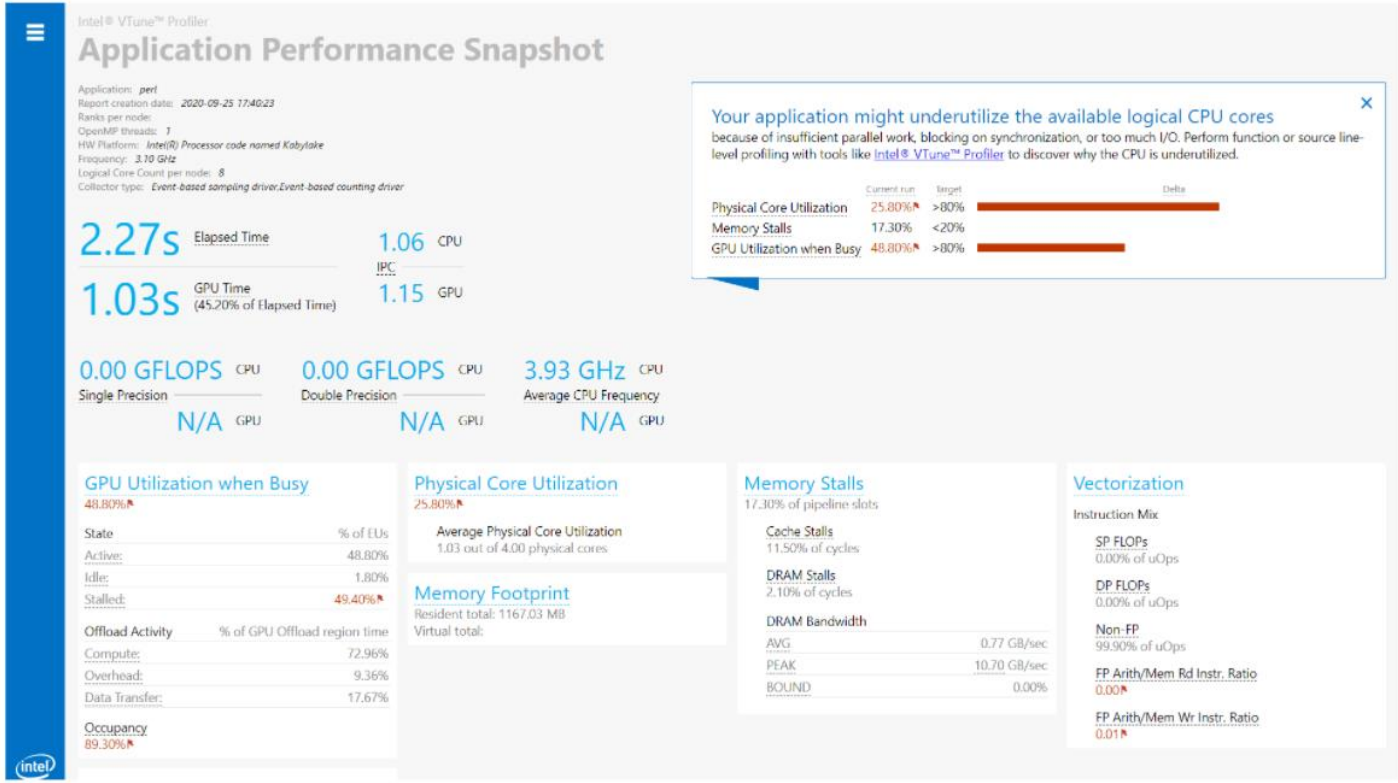
- “It’s like when the Internet came out and you found a business without a website. I’m starting to feel that way about an AI assistant. **How can they not have an assistant to help me?** It’s too manual. I want to offload the busy work.”

PN

Supply Chain Data Scientist

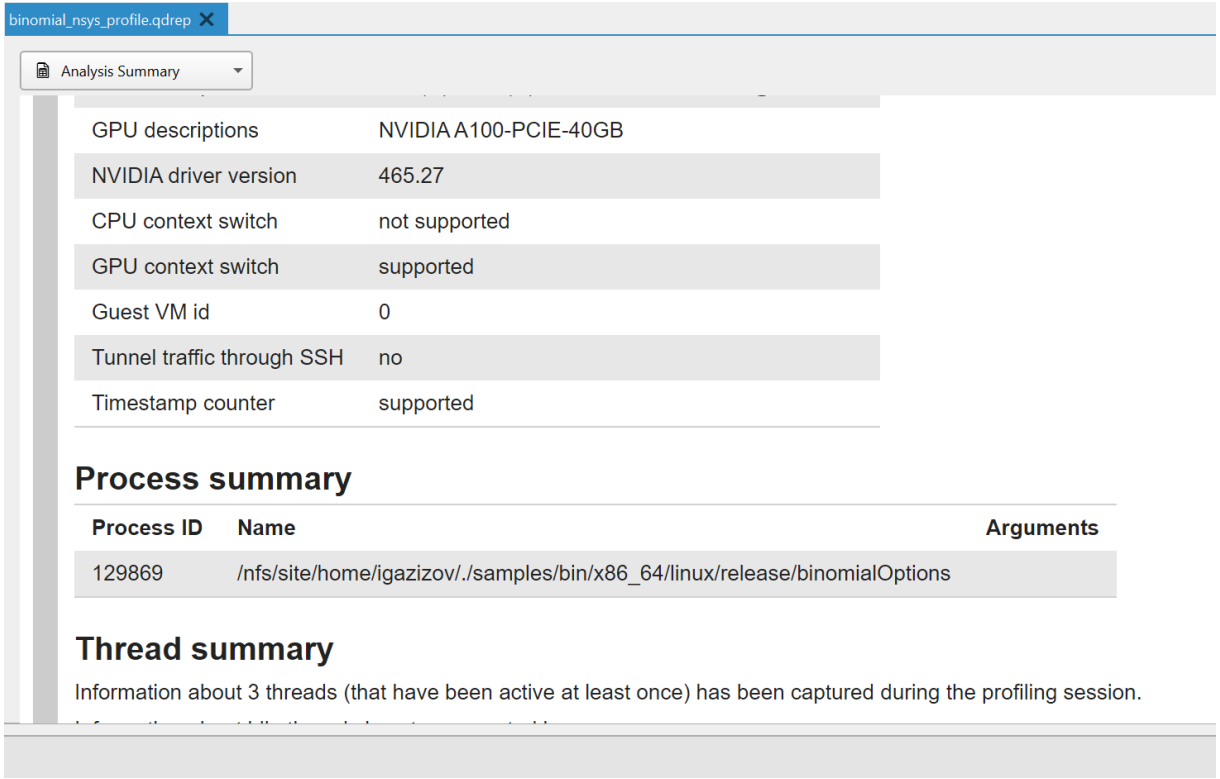
VTune: Performance Snapshot

A



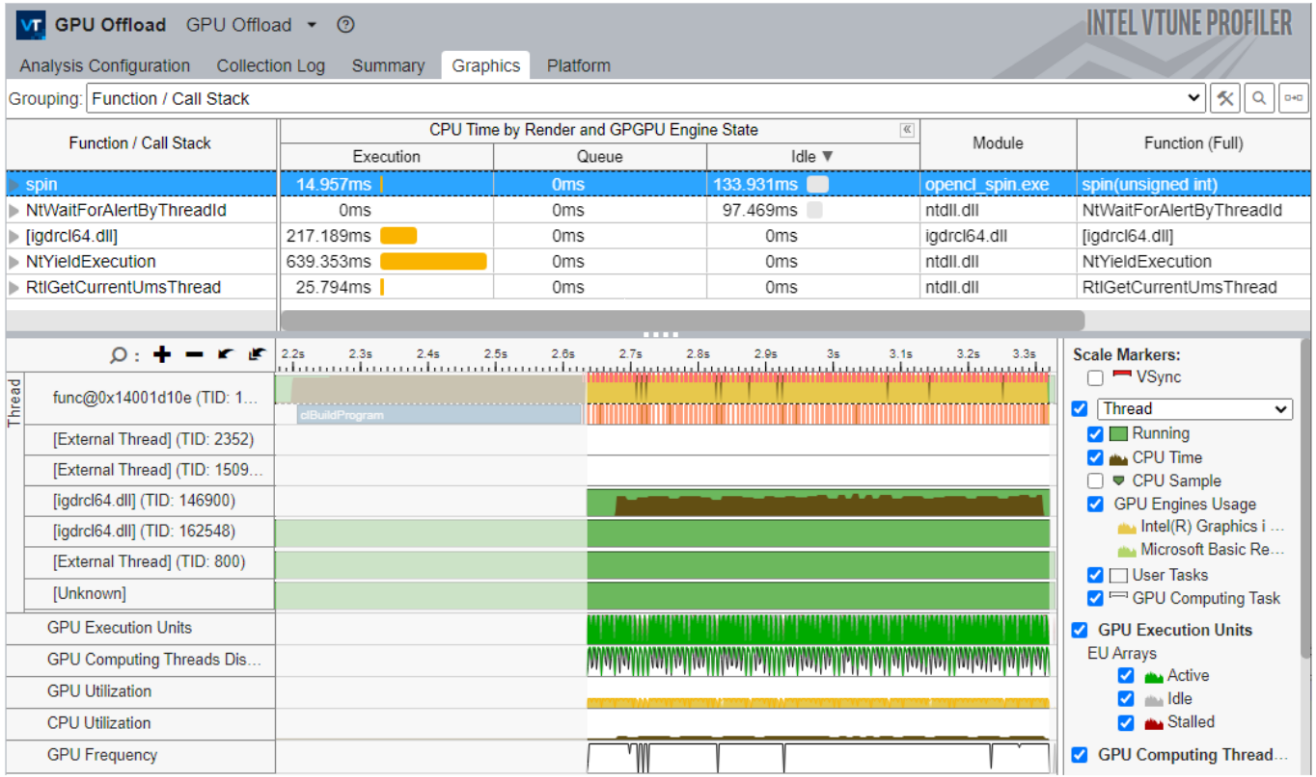
Nsight Systems: Analysis Summary

A



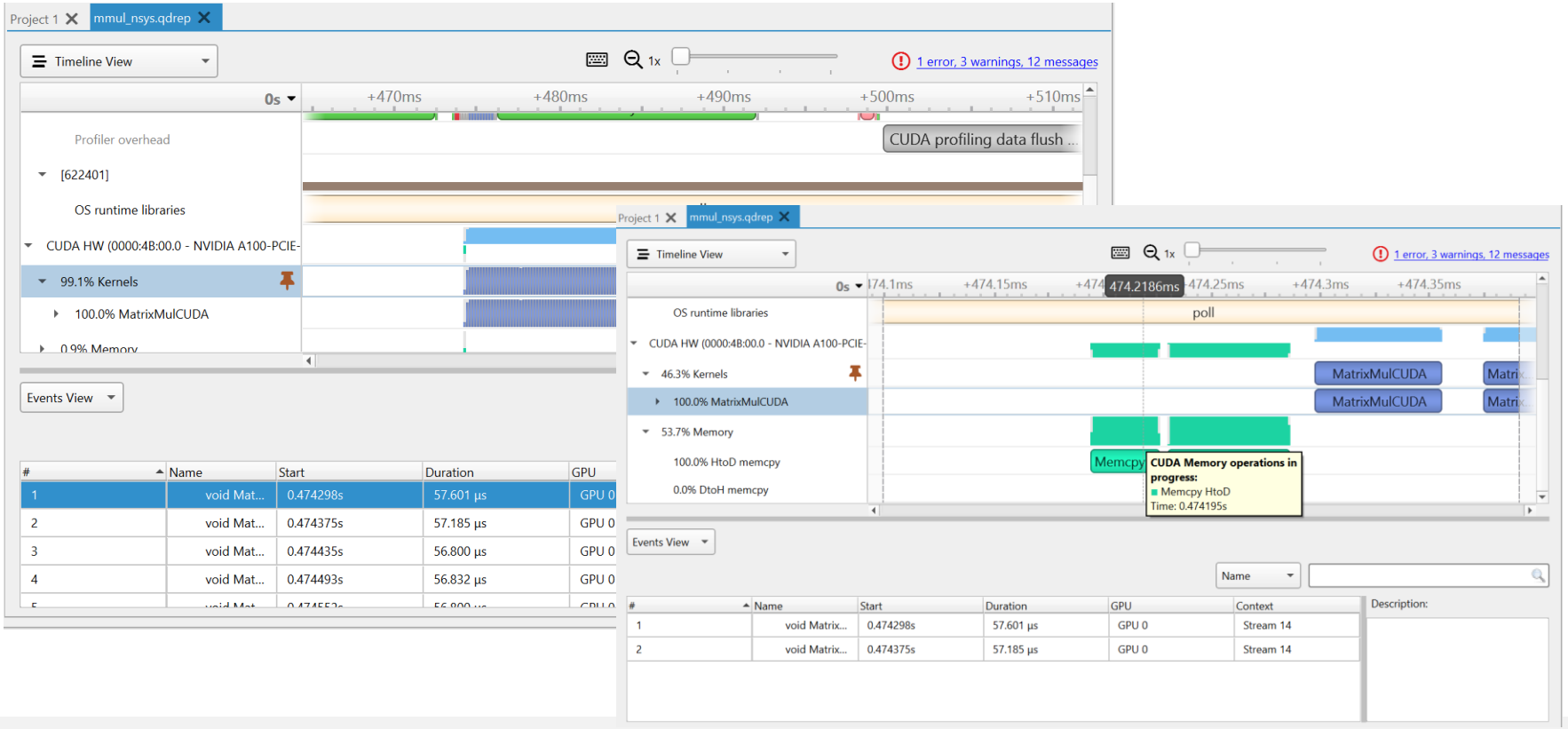
VTune: GPU Offload Analysis - Function/Call Stack Grouping

B

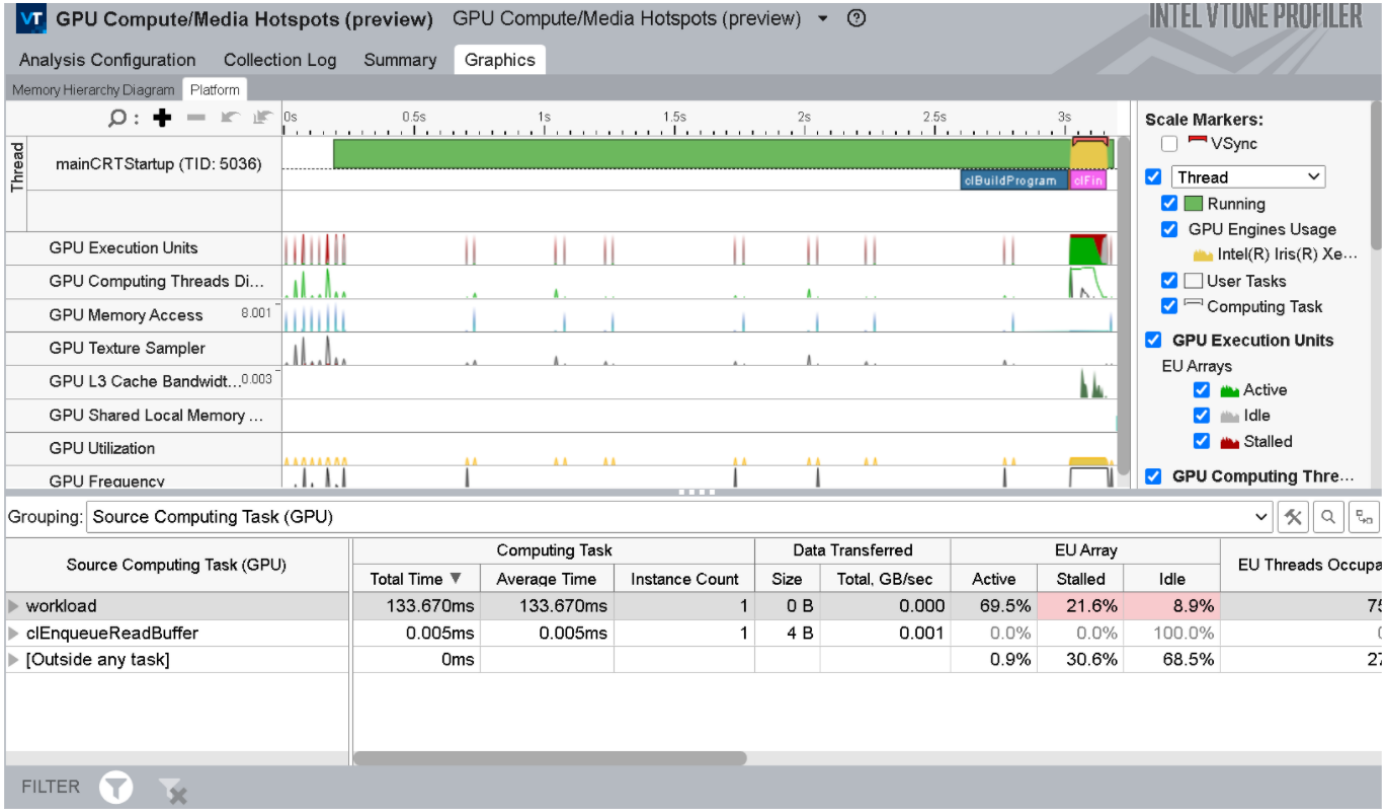


Nsight Systems: Timeline View

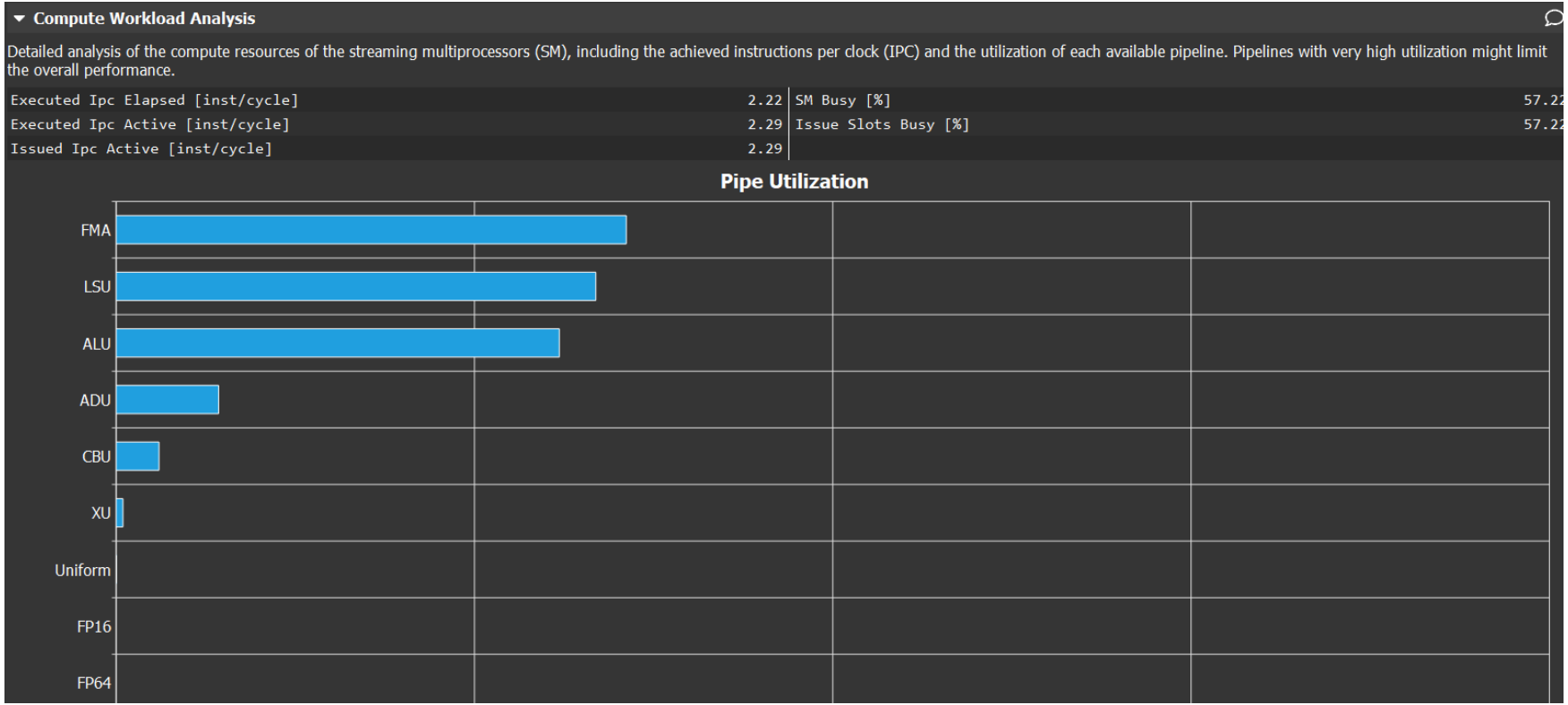
B



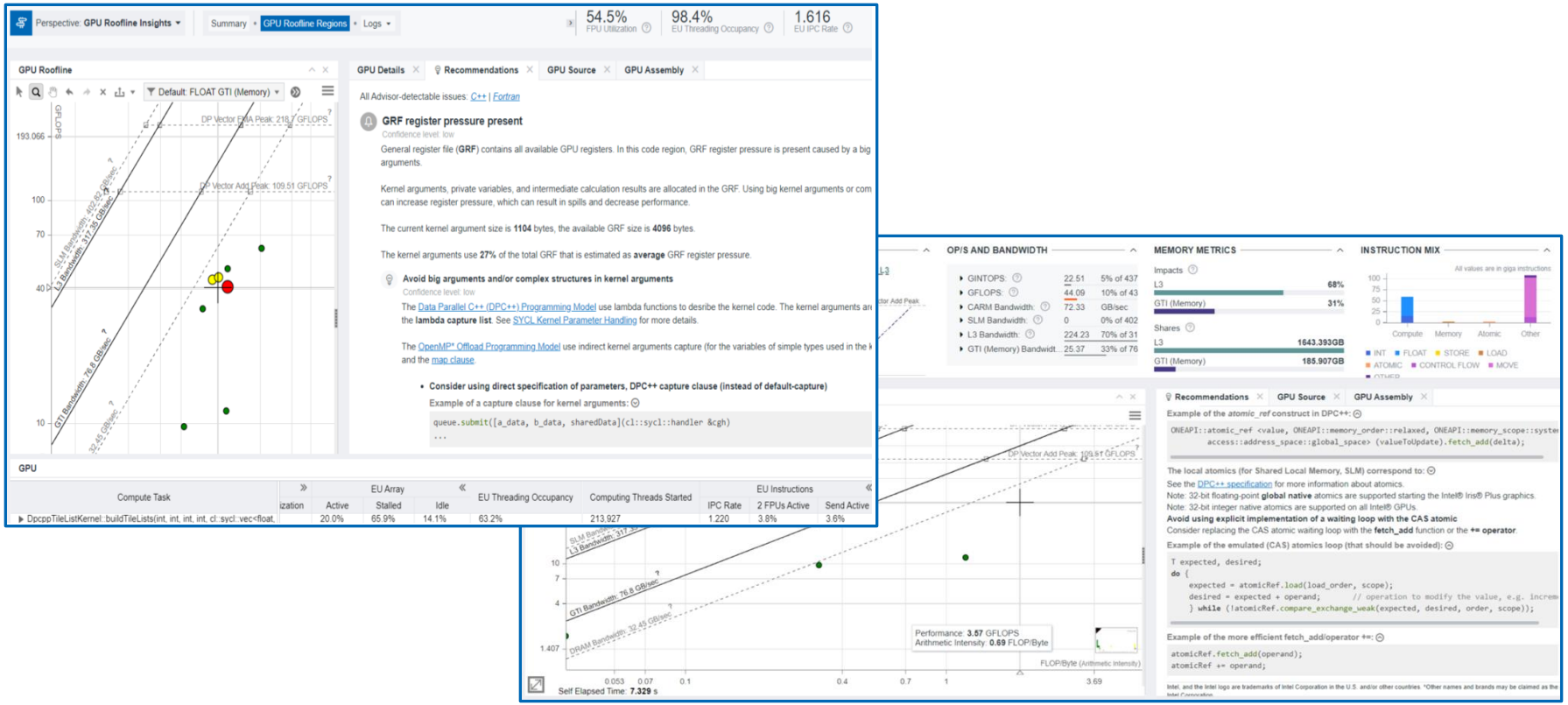
VTune: GPU Hotspot Analysis



Nsight Compute: Compute Workload Analysis

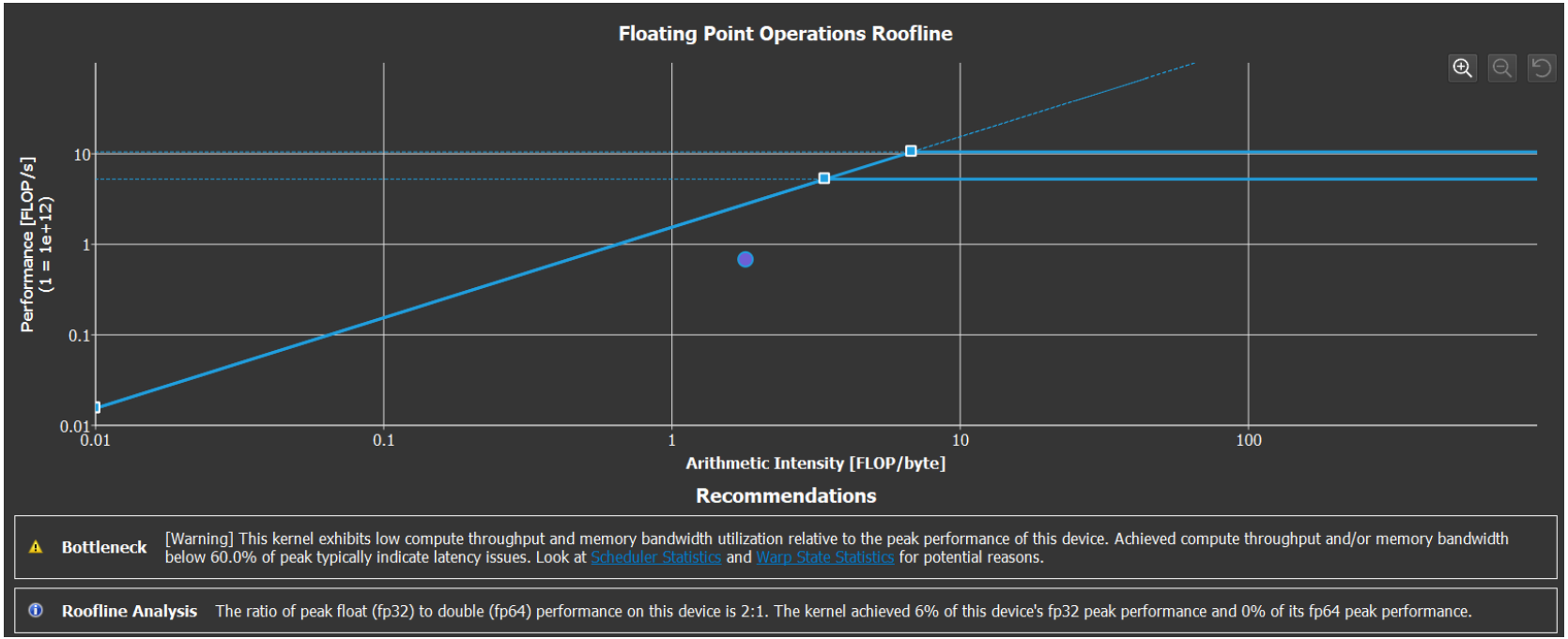


Intel Advisor: Roofline analysis

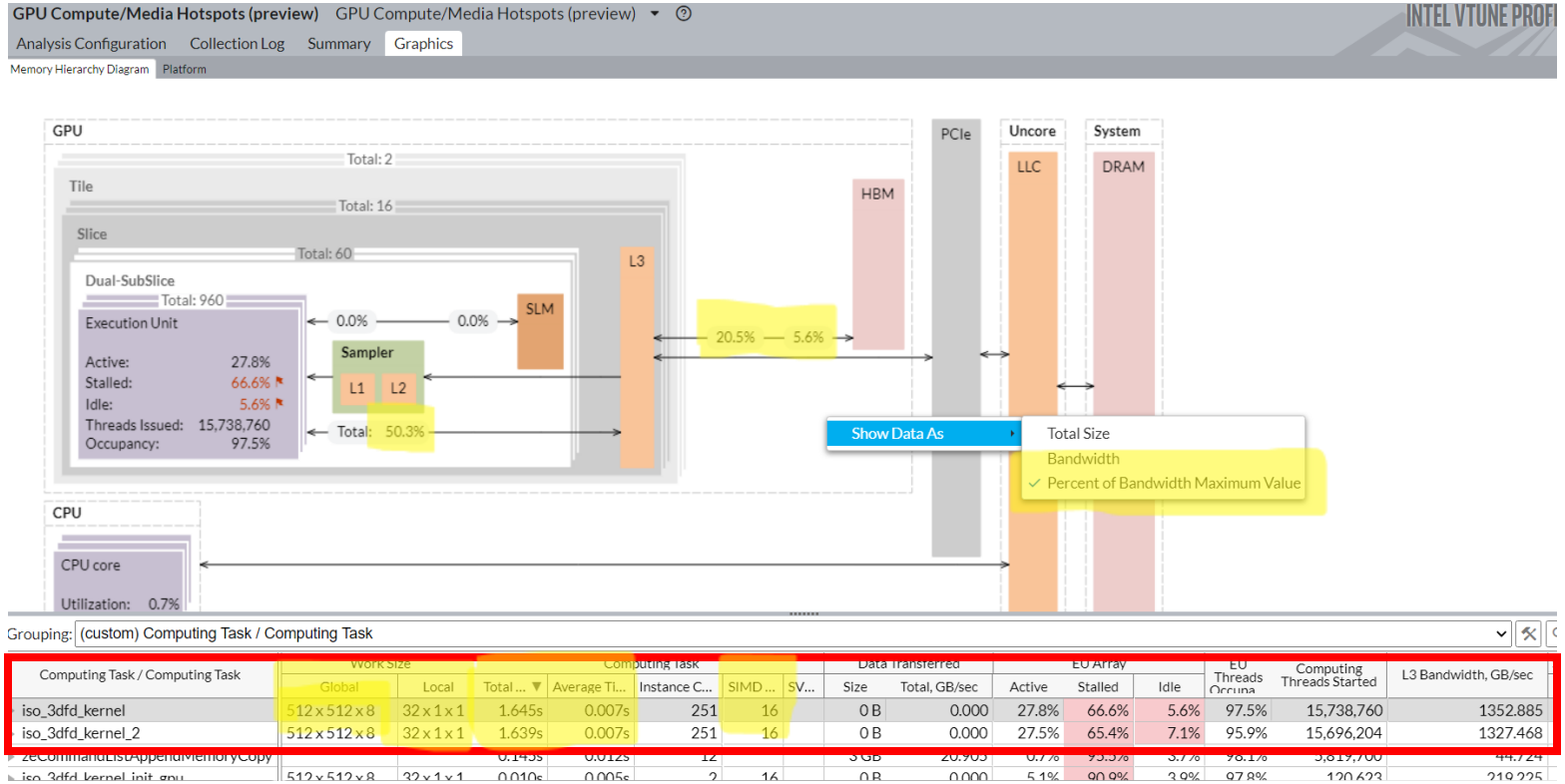


Nsight Compute: Roofline

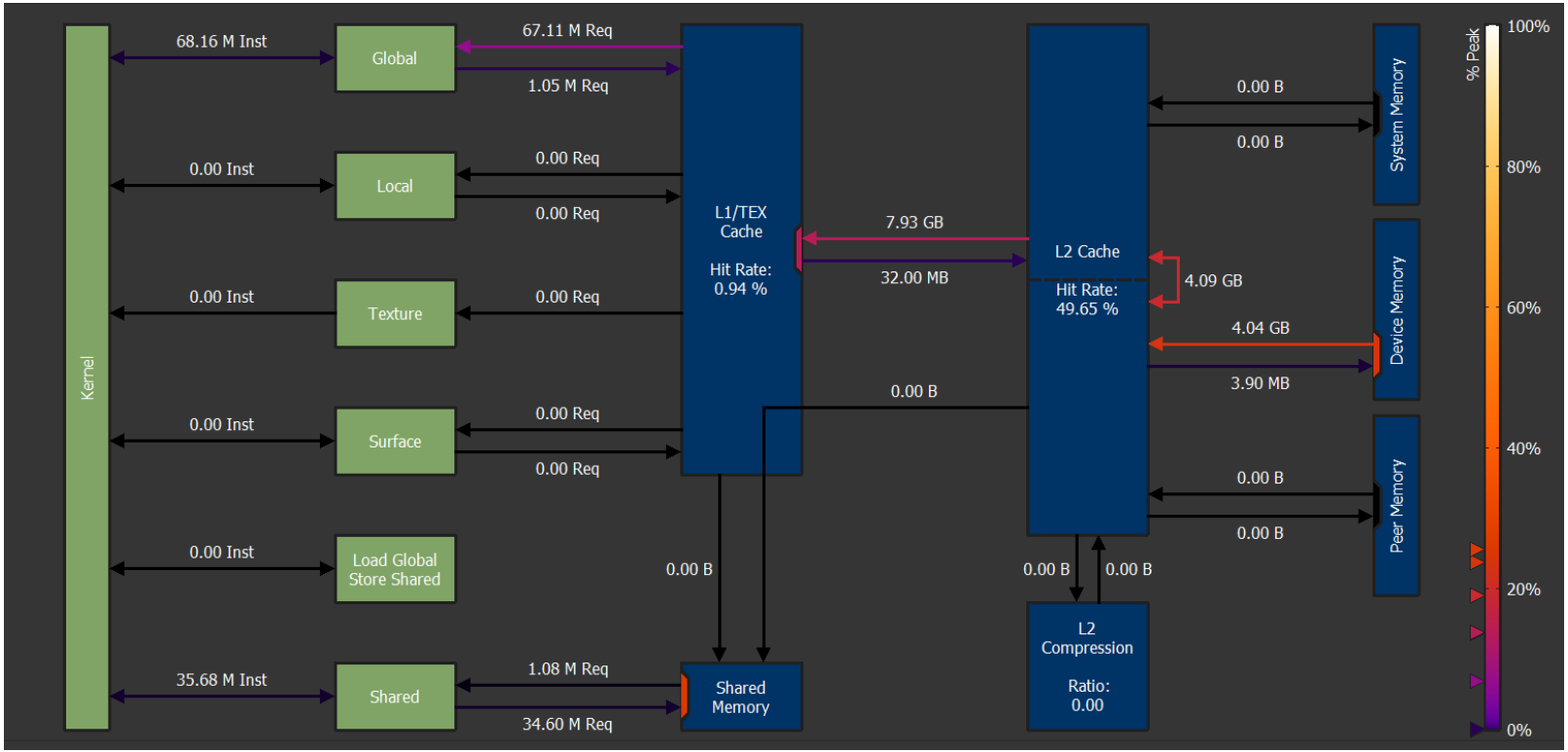
D



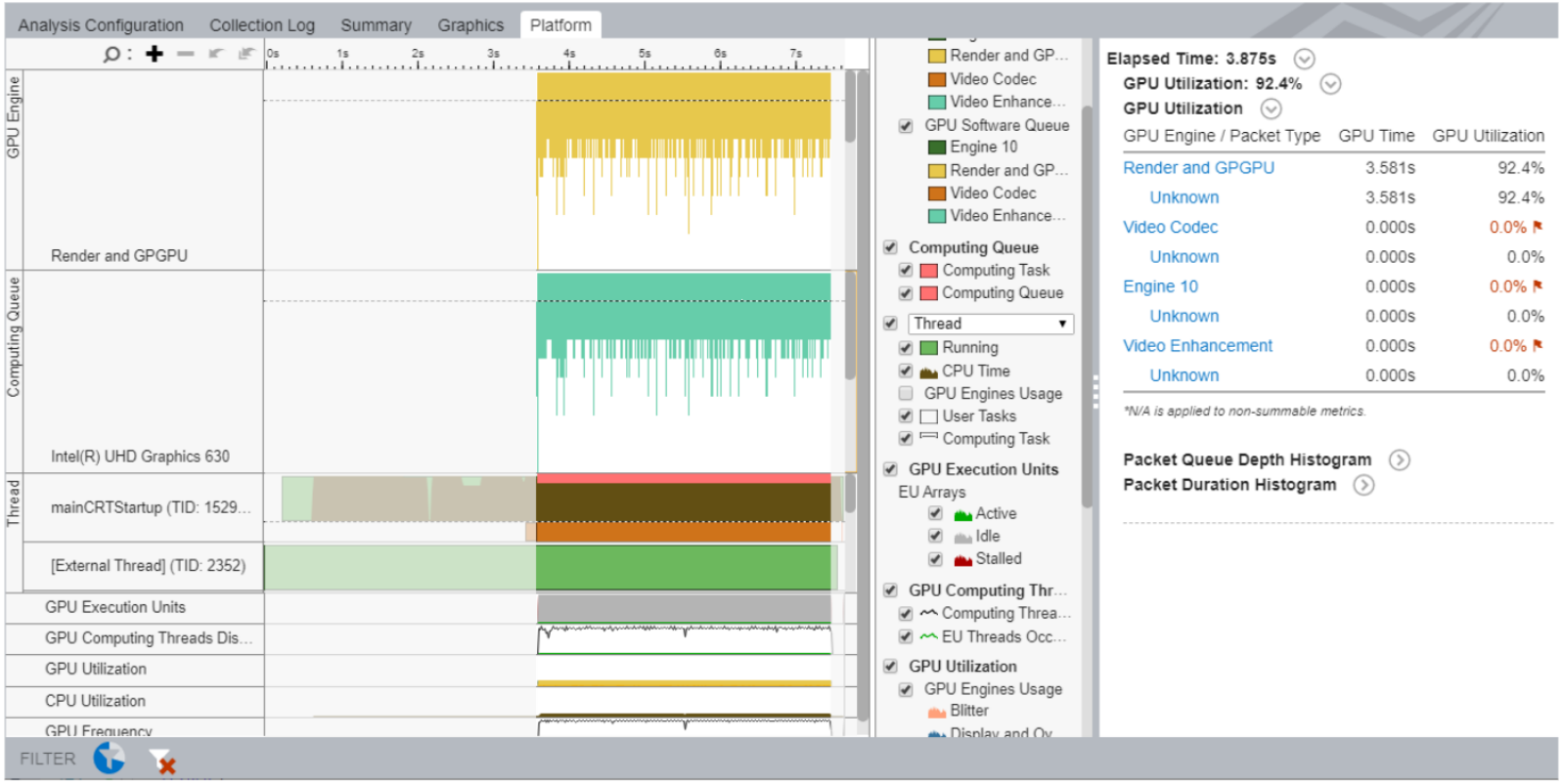
VTune: Memory Hierarchy Analysis



Nsight Compute: Memory Workload Analysis

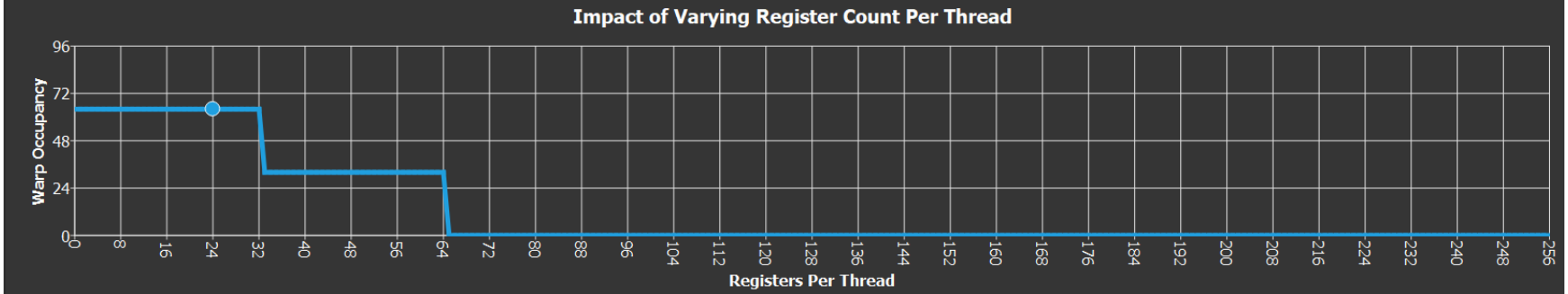


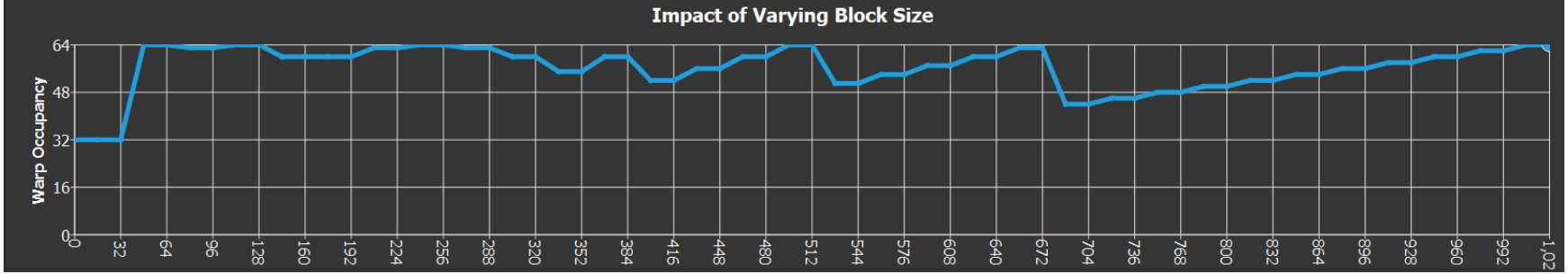
VTune: GPU Utilization Metrics



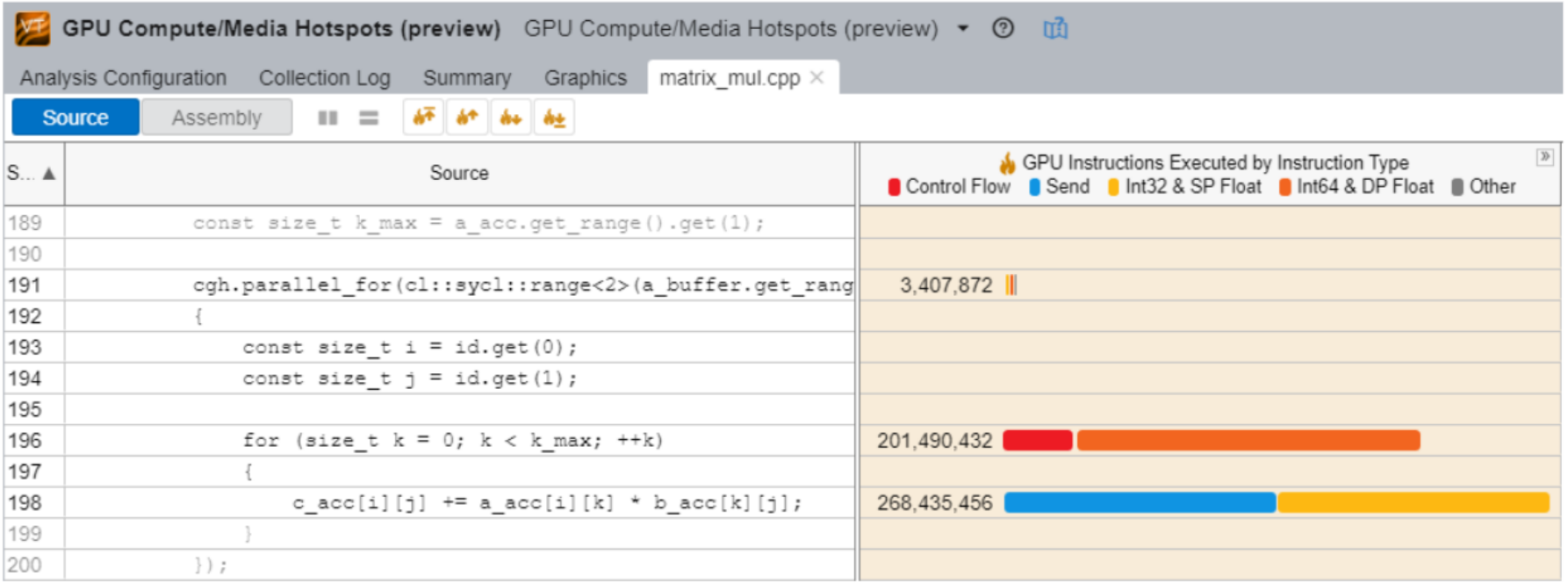
Nsight Compute: Occupancy

Theoretical Occupancy [%]	100	Block Limit Registers [block]	2
Theoretical Active Warps per SM [warp]	64	Block Limit Shared Mem [block]	32
Achieved Occupancy [%]	95.65	Block Limit Warps [block]	2
Achieved Active Warps Per SM [warp]	61.21	Block Limit SM [block]	32





VTune: Dynamic Instruction Count Analysis



Nsight Compute: Instruction Mix & Code Profile