Machine Learning

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

Predict GPU floating-point performance metrics based on hardware specifications

Jing Yang; jing.yang5@studio.unibo.it
Ge Li; ge.li2@studio.unibo.it

Catalogue

1 Proposal Analysis

2 Data Preprocessing

3 Modelling

4 Conclusion

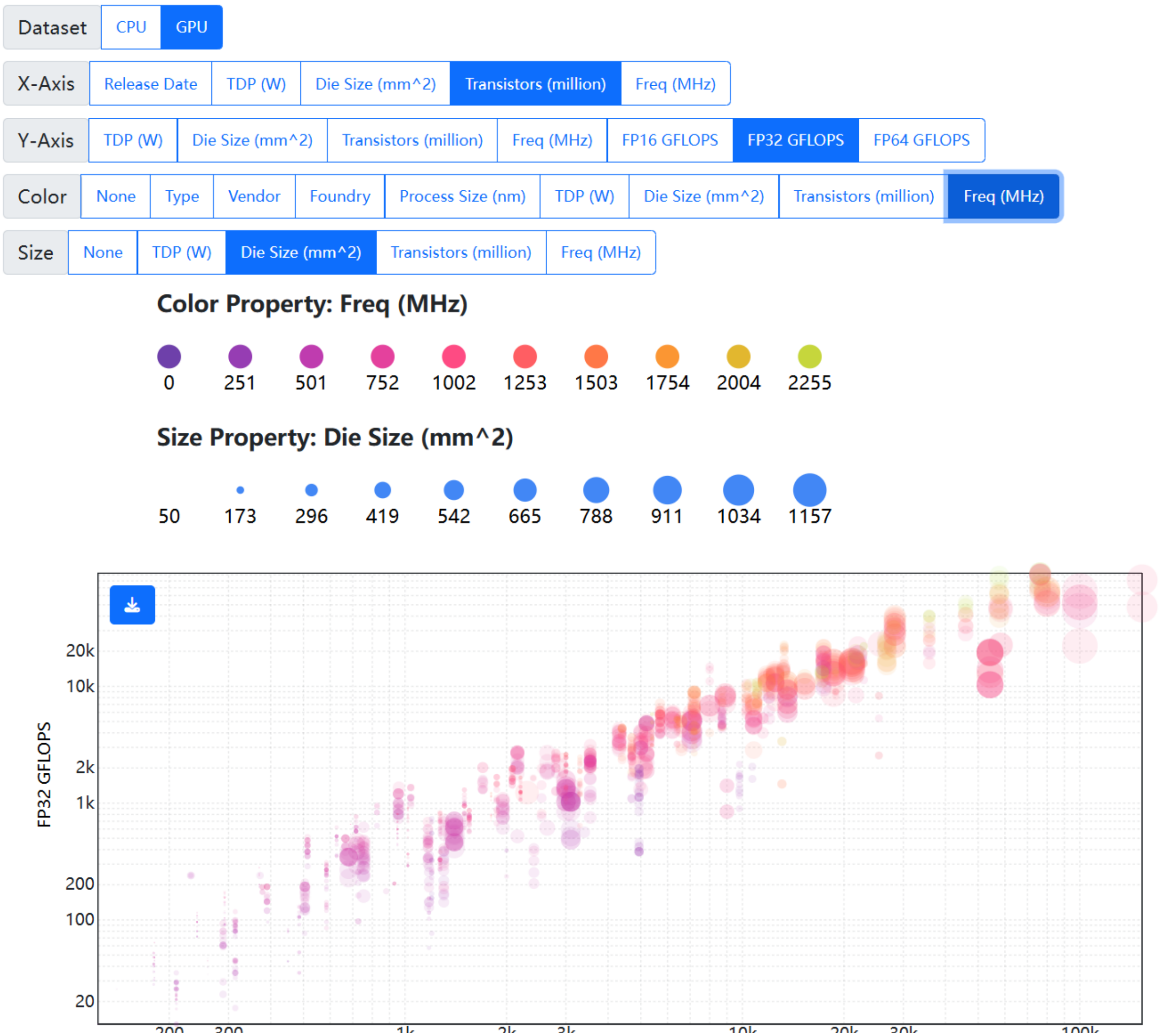
Proposal: Prediction of GPU performance

- **TDP (Thermal Design Power)**
- **Die Size**
- **Transistors**
- **Frequency**

GFLOPS:

Floating point operations per second (FLOPS, flops or flop/s) is a measure of computer performance in computing, there are different precision levels FP16,FP32 and FP64.

Related Works



Dataset CPU [✓ GPU]

X-Axis Release Date TDP (W) Die Size (mm^2) [✓ Transistors (million)] Freq (MHz)

Y-Axis TDP (W) Die Size (mm^2) Transistors (million) Freq (MHz) FP16 GFLOPS [✓ FP32 GFLOPS] FP64 GFLOPS

Color None Type Vendor Foundry Process Size (nm) TDP (W) Die Size (mm^2) Transistors (million) [✓ Freq (MHz)]

Size None TDP (W) [✓ Die Size (mm^2)] Transistors (million) Freq (MHz)

The original study*[Sun et al.(2019)]* aimed to investigate the validity of Moore's Law and Dennard Scaling amidst challenges posed by the practical limitations of transistor scaling.

By collecting data on over 4,000 CPUs and GPUs from vendors like Intel, AMD, and NVIDIA, the study identified that GPU performance improvements were significantly driven by increases in GPU frequency, improvements in the thermal design power (TDP), and growth in die size.

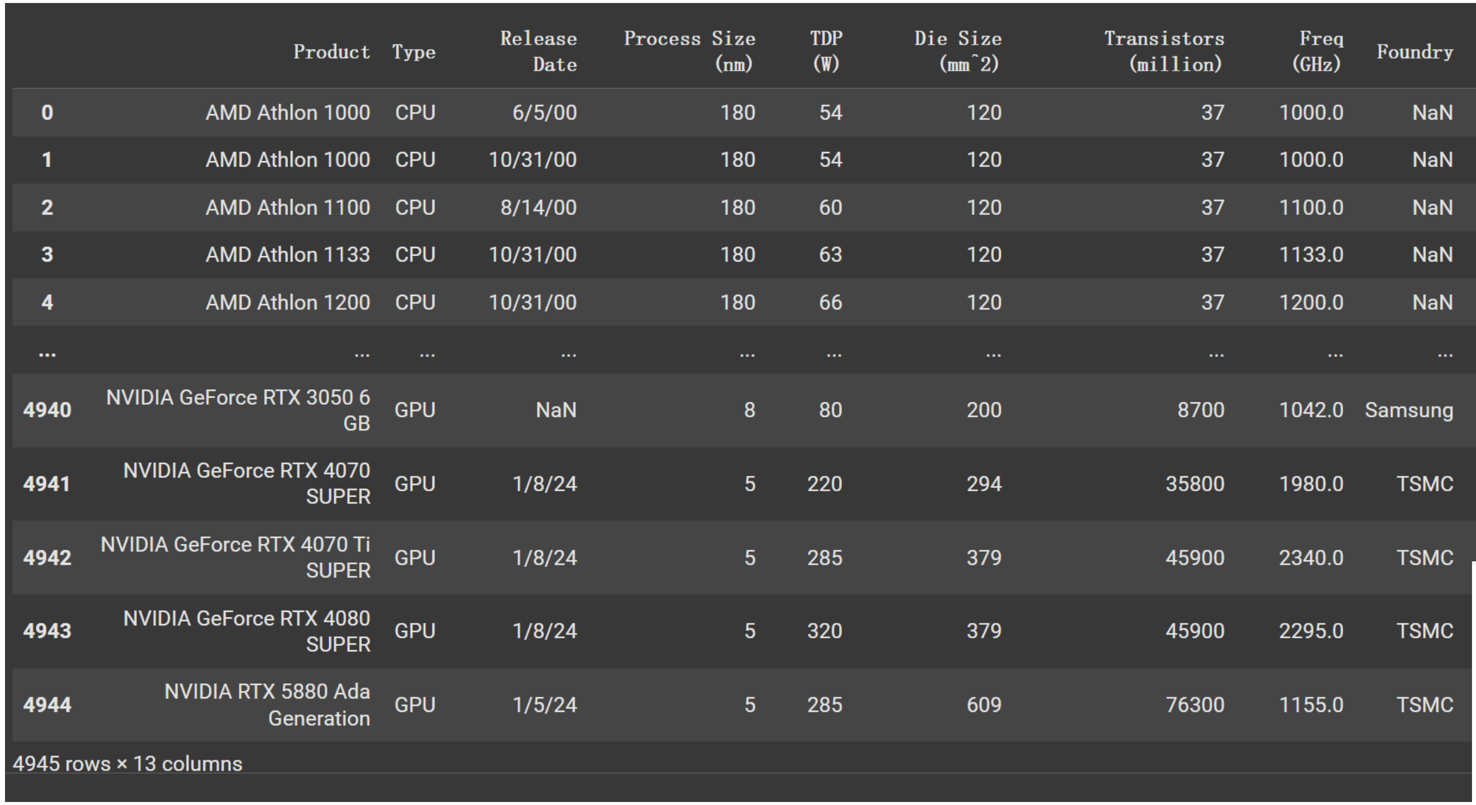
Another report by Epoch.ai* develops a log linear model, feature and model selection.

$$\log 10(\mathrm{FLOP/s}) = \alpha \cdot \log 10(X)$$

As a finding that almost all progress can be explained by two variables: transistor size and the number of cores.

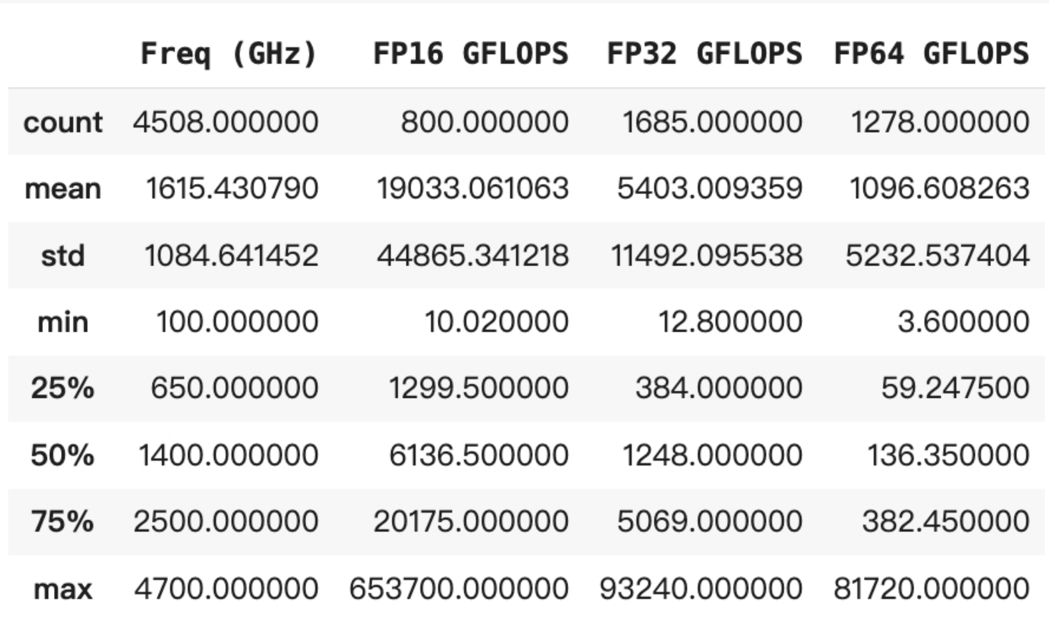
**https://epoch.ai/blog/predicting-gpu-performance#testing-non-linear-fits*

Data Understanding



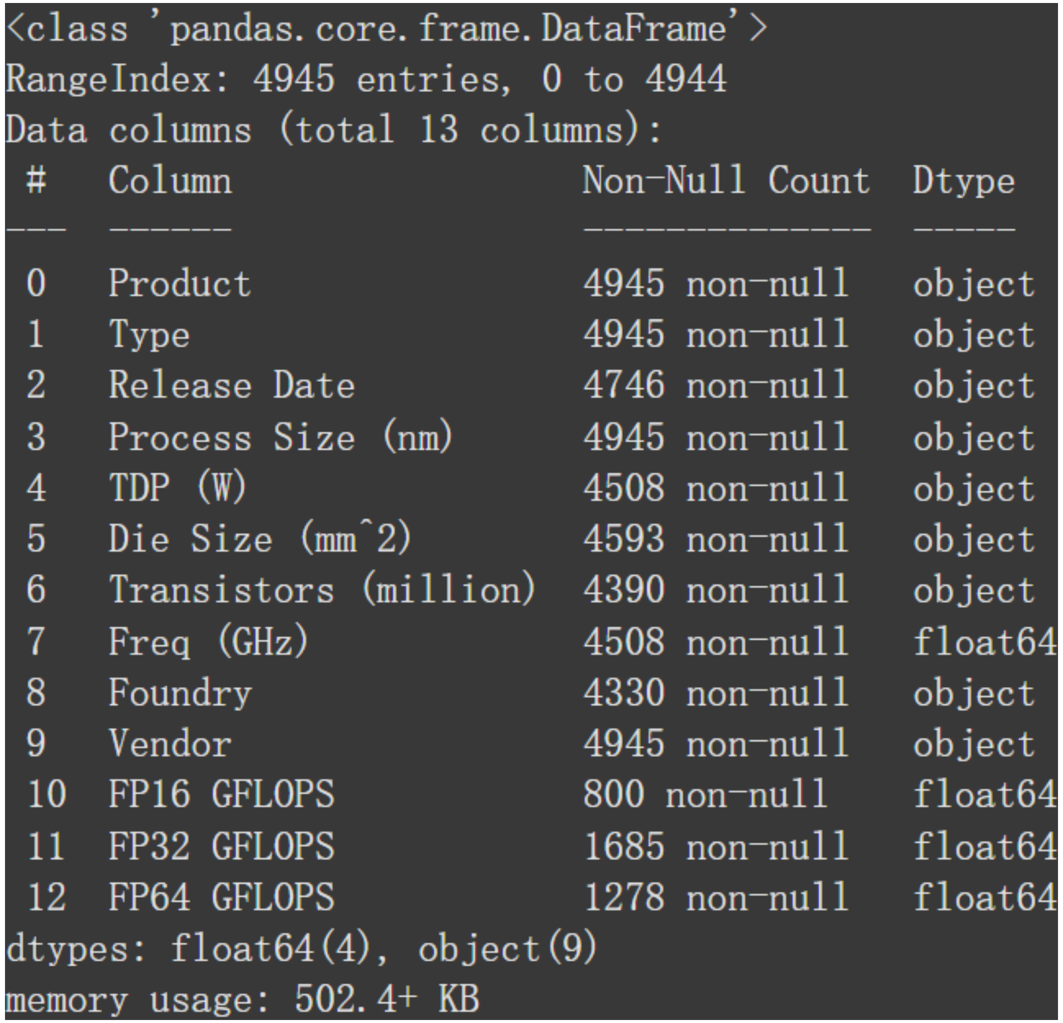
	Product	Type	Release Date	Process Size (nm)	TDP (W)	Die Size (mm^2)	Transistors (million)	Freq (GHz)	Foundry	Vendor	FP16 GFLOPS	FP32 GFLOPS	FP64 GFLOPS
0	AMD Athlon 1000	CPU	6/5/00	180	54	120	37	1000.0	NaN	AMD	NaN	NaN	NaN
1	AMD Athlon 1000	CPU	10/31/00	180	54	120	37	1000.0	NaN	AMD	NaN	NaN	NaN
2	AMD Athlon 1100	CPU	8/14/00	180	60	120	37	1100.0	NaN	AMD	NaN	NaN	NaN
3	AMD Athlon 1133	CPU	10/31/00	180	63	120	37	1133.0	NaN	AMD	NaN	NaN	NaN
4	AMD Athlon 1200	CPU	10/31/00	180	66	120	37	1200.0	NaN	AMD	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4940	NVIDIA GeForce RTX 3050 6 GB	GPU	NaN	8	80	200	8700	1042.0	Samsung	NVIDIA	6021.0	6021.0	94.08
4941	NVIDIA GeForce RTX 4070 SUPER	GPU	1/8/24	5	220	294	35800	1980.0	TSMC	NVIDIA	35480.0	35480.0	554.40
4942	NVIDIA GeForce RTX 4070 Ti SUPER	GPU	1/8/24	5	285	379	45900	2340.0	TSMC	NVIDIA	44100.0	44100.0	689.00
4943	NVIDIA GeForce RTX 4080 SUPER	GPU	1/8/24	5	320	379	45900	2295.0	TSMC				
4944	NVIDIA RTX 5880 Ada Generation	GPU	1/5/24	5	285	609	76300	1155.0	TSMC				

4945 rows × 13 columns



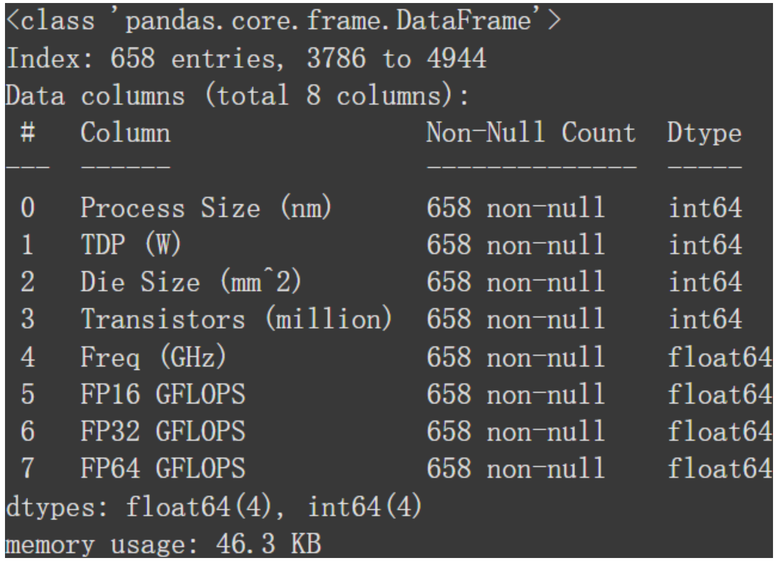
	Freq (GHz)	FP16 GFLOPS	FP32 GFLOPS	FP64 GFLOPS
count	4508.000000	800.000000	1685.000000	1278.000000
mean	1615.430790	19033.061063	5403.009359	1096.608263
std	1084.641452	44865.341218	11492.095538	5232.537404
min	100.000000	10.020000	12.800000	3.600000
25%	650.000000	1299.500000	384.000000	59.247500
50%	1400.000000	6136.500000	1248.000000	136.350000
75%	2500.000000	20175.000000	5069.000000	382.450000
max	4700.000000	653700.000000	93240.000000	81720.000000

Data Preprocessing



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4945 entries, 0 to 4944
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Product               4945 non-null   object
 1   Type                  4945 non-null   object
 2   Release Date          4746 non-null   object
 3   Process Size (nm)     4945 non-null   object
 4   TDP (W)               4508 non-null   object
 5   Die Size (mm^2)       4593 non-null   object
 6   Transistors (million) 4390 non-null   object
 7   Freq (GHz)            4508 non-null   float64
 8   Foundry               4330 non-null   object
 9   Vendor                4945 non-null   object
 10  FP16 GFLOPS           800 non-null    float64
 11  FP32 GFLOPS           1685 non-null   float64
 12  FP64 GFLOPS           1278 non-null   float64
dtypes: float64(4), object(9)
memory usage: 502.4+ KB
```

- Removing Non-GPU Data
- Removing Irrelevant Columns.
- Handling Missing and Unknown Values.
- Feature Engineering
- Data Type Conversion

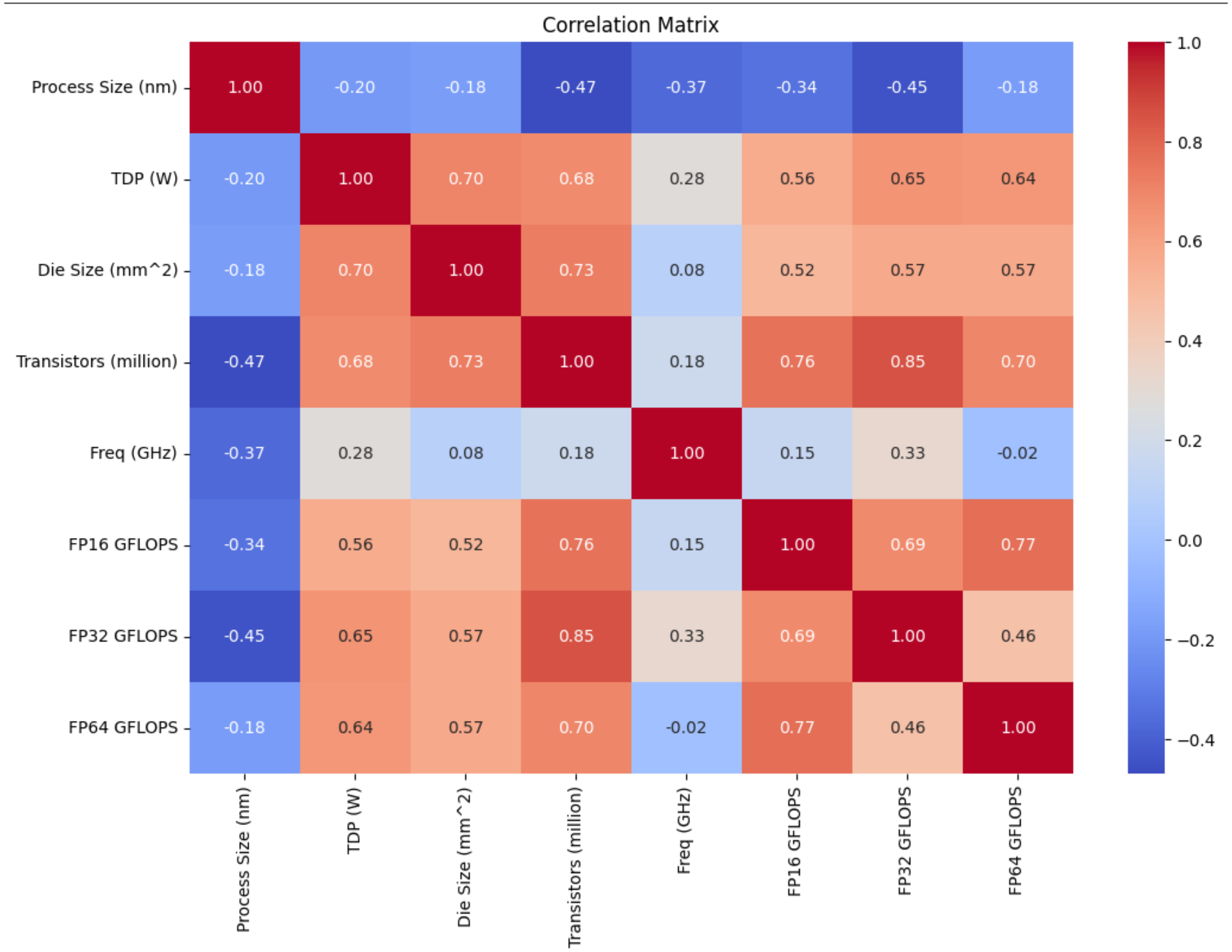


```
<class 'pandas.core.frame.DataFrame'>
Index: 658 entries, 3786 to 4944
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Process Size (nm)     658 non-null    int64
 1   TDP (W)               658 non-null    int64
 2   Die Size (mm^2)       658 non-null    int64
 3   Transistors (million) 658 non-null    int64
 4   Freq (GHz)            658 non-null    float64
 5   FP16 GFLOPS           658 non-null    float64
 6   FP32 GFLOPS           658 non-null    float64
 7   FP64 GFLOPS           658 non-null    float64
dtypes: float64(4), int64(4)
memory usage: 46.3 KB
```

Attributes Selection

- Selected features:
TDP, Die Size, Transistor Count.

- Rationale:
High correlation with target variables.

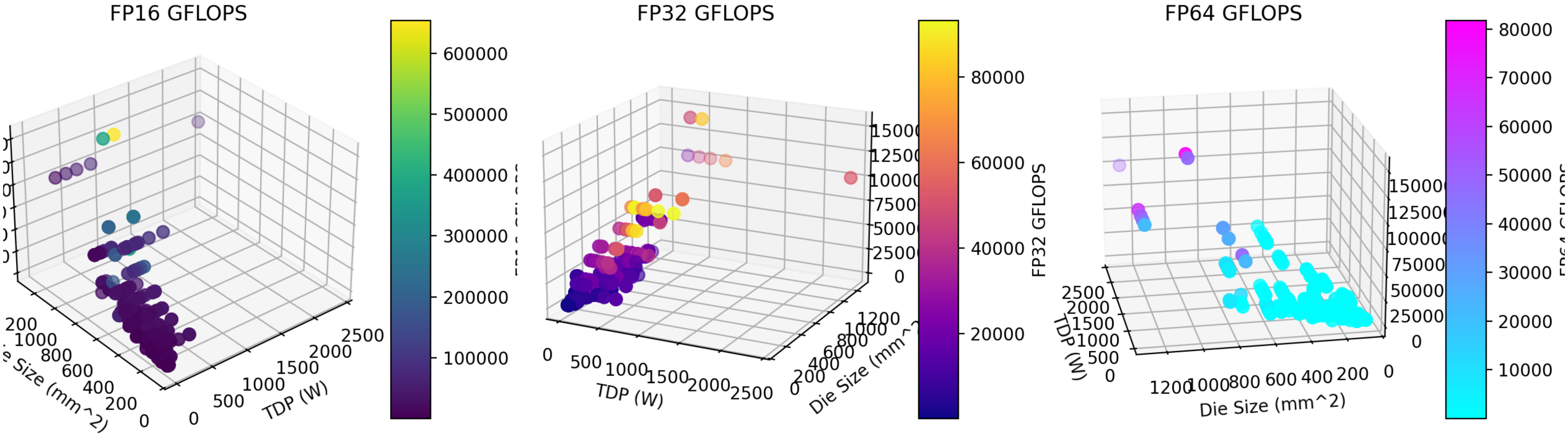


Correlation Matrix

	Process Size (nm)	TDP (W)	Die Size (mm^2)	Transistors (million)	Freq (GHz)	FP16 GFLOPS	FP32 GFLOPS	FP64 GFLOPS
Process Size (nm)	1.00	-0.20	-0.18	-0.47	-0.37	-0.34	-0.45	-0.18
TDP (W)	-0.20	1.00	0.70	0.68	0.28	0.56	0.65	0.64
Die Size (mm^2)	-0.18	0.70	1.00	0.73	0.08	0.52	0.57	0.57
Transistors (million)	-0.47	0.68	0.73	1.00	0.18	0.76	0.85	0.70
Freq (GHz)	-0.37	0.28	0.08	0.18	1.00	0.15	0.33	-0.02
FP16 GFLOPS	-0.34	0.56	0.52	0.76	0.15	1.00	0.69	0.77
FP32 GFLOPS	-0.45	0.65	0.57	0.85	0.33	0.69	1.00	0.46
FP64 GFLOPS	-0.18	0.64	0.57	0.70	-0.02	0.77	0.46	1.00

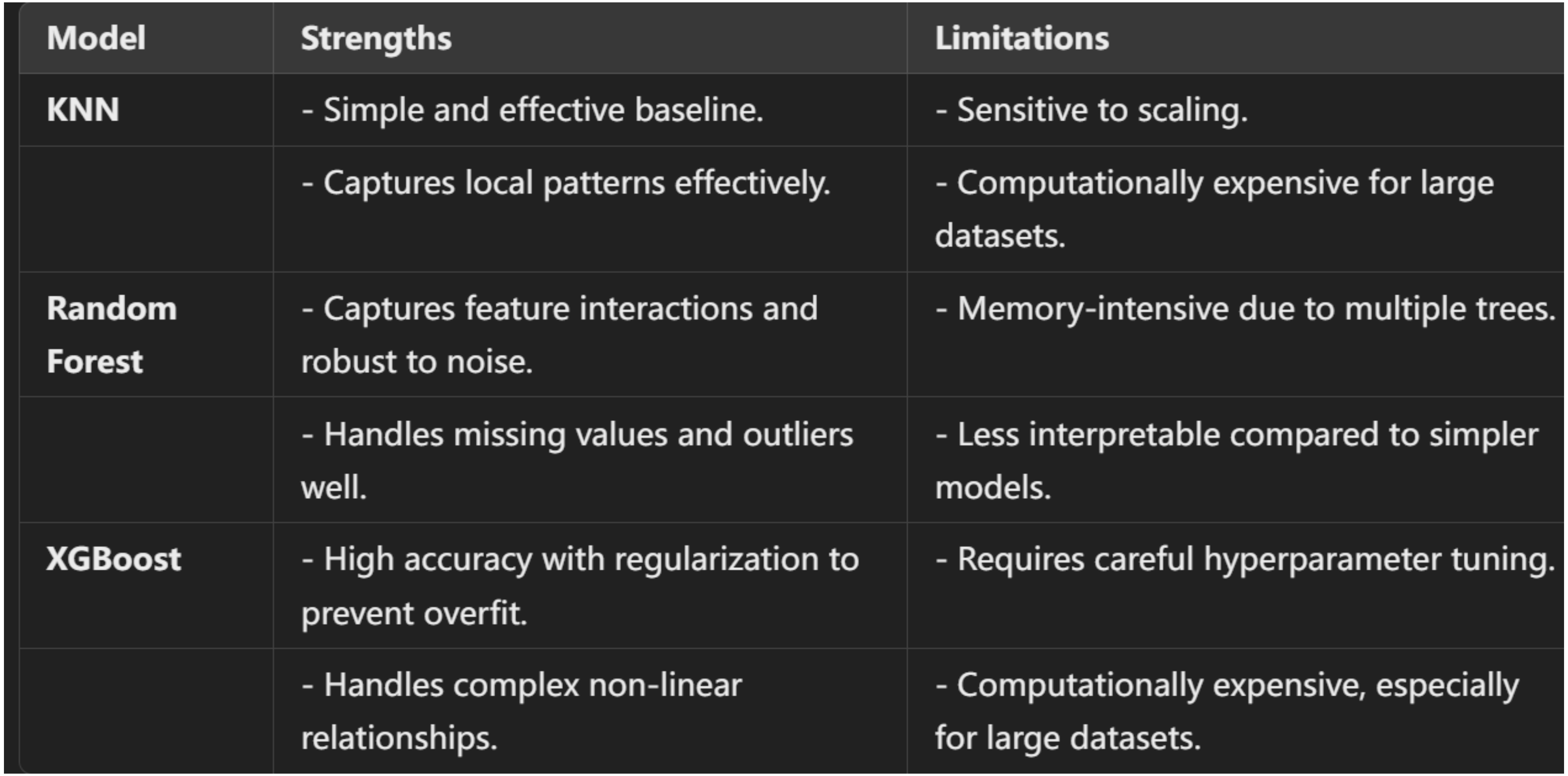
Data Visualization

3D Visualization of GPU Parameters for FP16, FP32, and FP64 GFLOPS



How GPU performance(in different floating point precision) reflects on those features, which indicates the diversification of the tasks?

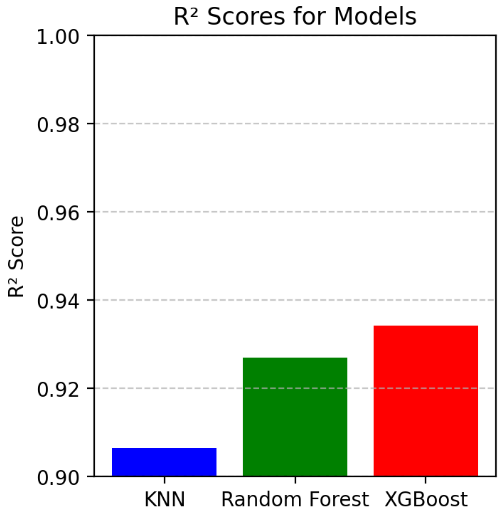
Model Selection

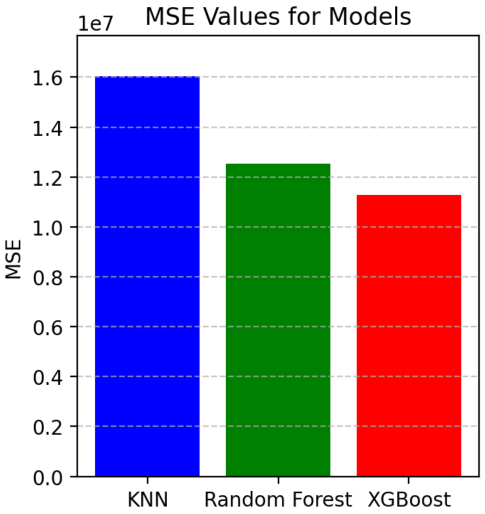


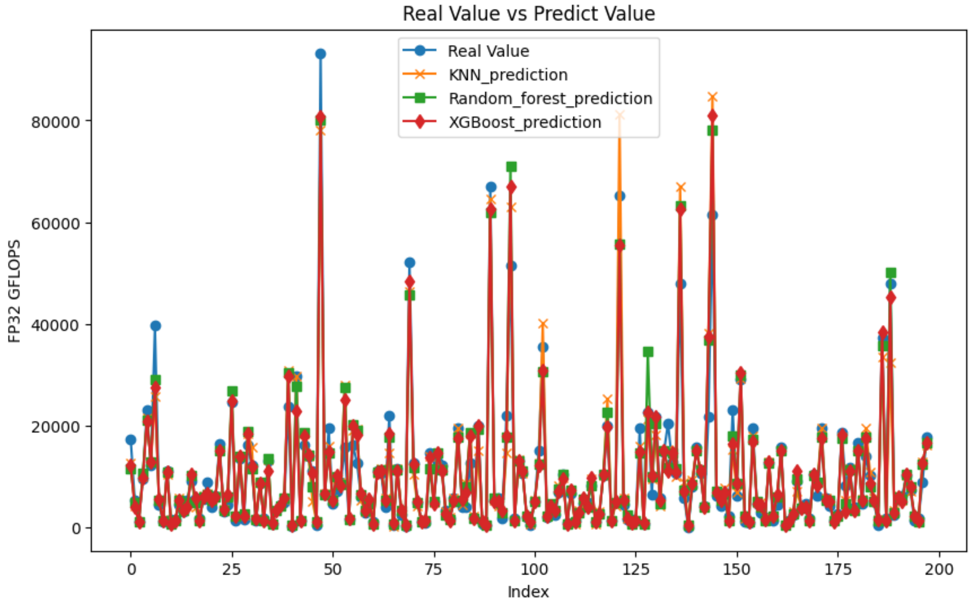
Model	Strengths	Limitations
KNN	- Simple and effective baseline.	- Sensitive to scaling.
	- Captures local patterns effectively.	- Computationally expensive for large datasets.
Random Forest	- Captures feature interactions and robust to noise.	- Memory-intensive due to multiple trees.
	- Handles missing values and outliers well.	- Less interpretable compared to simpler models.
XGBoost	- High accuracy with regularization to prevent overfit.	- Requires careful hyperparameter tuning.
	- Handles complex non-linear relationships.	- Computationally expensive, especially for large datasets.

Results

	KNN	Random Forest	Xgboost
R2	0.9064	0.9269	0.9342
MSE	16,046,345	12,535,500	11,288,906





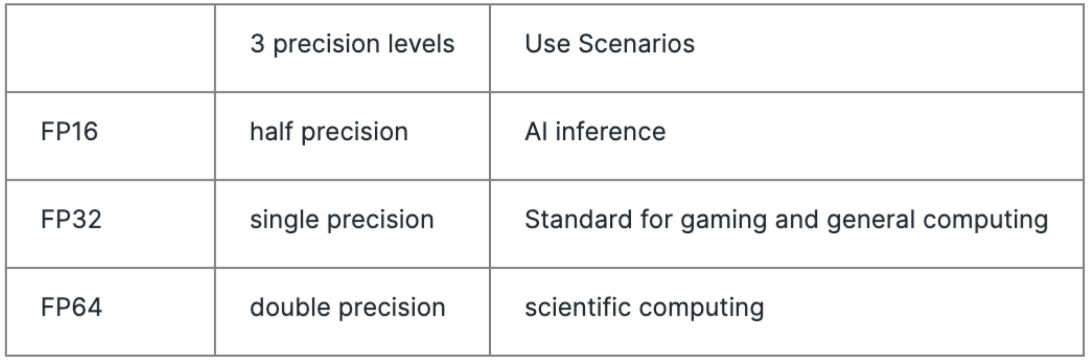


Conclusion&What we learnt

1. We used the updated dataset which includes records until Jan 2024 and focuses on using machine learning to predict GPU floating-point performance metrics (FP32 GFLOPS) based on hardware specifications.

2. We create a hierarchy of parameter importance to see which factor impact GPU performance the most. Transistor Count is shown as the dominant predictor of GPU performance.

3. We developed a 3D visualization where we make the features and the targeted variable on the same plot to see the feature interactions.



	3 precision levels	Use Scenarios
FP16	half precision	AI inference
FP32	single precision	Standard for gaming and general computing
FP64	double precision	scientific computing

« The design philosophy NVIDIA seems to use is to build a base configuration with full FP32 performance (which is needed for 3D graphics, their bread & butter business), then bolt on additional units (FP16, FP64) for professional markets, where the additional revenue per part (thousands of dollars) more than makes up for higher die costs (hundreds of dollars).

This approach allows their consumer line to compete on price, while allowing their professional line to compete on performance and features. » —CUDA forum

Reference

[Dally et al.(2021)] William J Dally, Stephen W Keckler, and David B Kirk. 2021. **Evolution of the graphics processing unit (GPU)**. IEEE Micro 41, 6 (2021), 42–51.

[Dongarra et al.(2024)] Jack Dongarra, John Gunnels, Harun Bayraktar, Azzam Haidar, and Dan Ernst. 2024. **Hardware Trends Impacting Floating-Point Computations In Scientific Applications.** arXiv preprint arXiv:2411.12090 (2024).

[Sun et al.(2019)] Yifan Sun, Nicolas Bohm Agostini, Shi Dong, and David Kaeli. 2019. **Summarizing CPU and GPU design trends with product data.** arXiv preprint arXiv:1911.11313 (2019).

Predicting GPU Performance. Report from Epoch AI.
https://epoch.ai/blog/predicting-gpu-performanceappendix-c—physical-size-limits-of-transistors-details