Lesertest: Samsung 9100 Pro – Der Speicher-Turbo für lokale KI und Virtualisierung?

Gen 5 vs. Gen 4 vs. SATA: Ein Praxis-Check jenseits von CrystalDiskMark

Inhalt

1. Einleitung und Motivation

Speicherplatz ist heutzutage günstig, doch Speicher-*Geschwindigkeit* bleibt ein Luxusgut. Wer seinen PC nur für Office-Anwendungen oder Spiele nutzt, spürt zwischen einer guten PCIe 3.0 SSD und einem modernen PCIe 5.0 Modell oft kaum einen Unterschied. Doch wie sieht es aus, wenn der Rechner zur Workstation wird?

Als jemand, der sich gerne mit **lokaler Künstlicher Intelligenz (LLMs, Stable Diffusion)**, **Software-Entwicklung** und **Virtualisierung** beschäftigt, stelle ich mir eine entscheidende Frage:

"Lohnt sich der Aufpreis für die theoretische Bandbreite von PCIe 5.0 in der Praxis wirklich, oder wartet meine Grafikkarte am Ende doch nur auf die CPU?"

Synthetische Benchmarks wie CrystalDiskMark liefern zwar beeindruckende Zahlen, sagen aber nicht alles über das Verhalten beim Laden eines 8-Milliarden-Parameter-Sprachmodells oder beim Starten eines komplexen Docker-Containers aus. Mit diesem Lesertest der **Samsung 9100 Pro (2 TB)** möchte ich genau diese Lücke schließen. Mein Ziel ist es, nicht nur Balken zu vergleichen, sondern herauszufinden, ob und wo eine High-End-SSD im Creator-Alltag echte Lebenszeit spart.

2. Das Testfeld: David gegen Goliath

Um die Leistung der Samsung 9100 Pro fair einzuordnen, tritt sie nicht im luftleeren Raum an. Sie muss sich gegen eine breite Palette an Speichertechnologien behaupten, die stellvertretend für typische Aufrüst-Szenarien stehen:

- **Der Herausforderer (Gen 5):** Samsung 9100 Pro (2 TB) – *Das Testmuster*
- **Die Referenz (Gen 5):** Crucial T705 (1 TB) – *Der aktuelle "Speed-King"*
- **Der Vernunft-Standard (Gen 4):** Samsung 990 Pro (2 TB) – *Beliebt und bewährt*
- **Die Veteranen (SATA):** Crucial BX100 (250 GB) & OCZ Vertex 3 (120 GB) – *Ältere Modelle mit DRAM-Cache*
- **SATA Entry:** SanDisk Plus (120 GB) – *Günstiges Modell ohne DRAM-Cache*

3. Methodik: Reproduzierbarkeit statt Zufall

Wer SSDs unter Windows testet, kämpft gegen einen unsichtbaren Gegner: Das Betriebssystem selbst. Windows nutzt freien Arbeitsspeicher aggressiv als Datei-Cache (Standby List). Startet man eine Anwendung oder lädt ein KI-Modell zum zweiten Mal, kommen die Daten oft gar nicht mehr von der SSD, sondern direkt aus dem RAM. Das

Ergebnis wären traumhafte, aber völlig unrealistische Messwerte, die nichts mit der Leistung des Laufwerks zu tun haben.

Um echte "Real-World-Performance" zu messen, habe ich mich nicht auf manuelle Stoppuhren verlassen, sondern ein vollautomatisiertes Benchmark-Framework (PowerShell & Python) entwickelt – auf Basis dessen wurden die Versionen 1 und 2 dieses Testes veröffentlicht.

Leider hat dieses Framework nach genauem Review bewiesen, dass es für mich nicht erklärliche Diskrepanzen bei den Messwerten zwischen manuellen und automatisierten Messungen gab. Deswegen habe ich nun für Version 3 des Berichts auf manuelle Auswertungen zurückgegriffen. Die Formulierung des Berichts kommt aber weiterhin in weiten Teilen von Gemini 3 Pro.

3.1. **Der "Cold Start" Enforcer**

Auch für die manuelle Auswertung der SSD Ergebnisse bleibt die Kontrolle über den Windows-Cache sehr wichtig. Vor jedem kritischen Messdurchlauf (gekennzeichnet als "Cold Run") habe ich mit dem Tool RamMap (Sysinternals) die Standby List geleert.

Funktionen und Effekt:

- Funktion: Es zwingt Windows, den Datei-Cache im RAM sofort zu leeren.
- Effekt: Die SSD muss jedes Byte physisch neu lesen. Nur so lässt sich unterscheiden, ob eine SSD Daten mit 500 MB/s (SATA) oder 10.000 MB/s (Gen 5) liefert.

3.2. **Die "Junction-Mapping" Technik**

KI-Tools wie Ollama erwarten ihre Modelle in festen Verzeichnissen (z.B. %USERPROFILE%\.ollama\models). Ein einfaches "Installieren auf Laufwerk D:" ist oft umständlich oder verändert das Verhalten der Software.

Mein Framework nutzt daher NTFS Junctions (Verknüpfungen):

1. Die großen Modell-Daten (LLaMA 3, SDXL Checkpoints) liegen physisch auf der jeweiligen Test-SSD (z.B. Laufwerk E:).
2. Vor jedem Test habe ich manuell eine Verknüpfung vom Standard-Pfad C:\Users\...\models auf den jeweiligen Speicherort der Modelle gelegt.
3. Für die Software sieht es so aus, als lägen die Daten am gewohnten Ort .

Vorteil: Das Testsystem (OS, Treiber, Software-Versionen) bleibt für jede SSD zu 100% identisch. Nur das physische Speichermedium unter der Haube wird ausgetauscht.

3.3. **Die Test-Szenarien im Detail**

Um ein wirklich umfassendes Bild zu zeichnen, durchläuft jede SSD sechs Disziplinen, die vom reinen Konsumieren (Inference) über das Entwickeln (Dev-Ops) bis hin zum harten Dateitransfer reichen:

Synthetische Basis: Klassische Messungen mit *CrystalDiskMark* um die theoretische Maximalleistung und Vergleichbarkeit mit anderen Reviews sicherzustellen.

Inference & VRAM-Loading (Ollama): Gemessen wird die Zeit, bis ein LLM (Large Language Model) von der SSD gelesen, deserialisiert und in den Grafikspeicher (VRAM) geladen ist.

Modelle: LLaMA 3 (8B), Mistral (7B), Phi-3 Mini.

Metrik: "Time to First Token" (Ladezeit).

Content Creation (ComfyUI / Stable Diffusion): Starten einer komplexen Python-Umgebung inkl. Laden eines 6 GB großen Checkpoints.

Developer Workflows (Hugging Face & Datasets): Hier simulieren wir den Alltag eines KI-Entwicklers, was sich mit dem Tool fio (https://github.com/axboe/fio) sehr leicht bewerkstelligen lässt.

Hugging Face Cache Load: Um einen Hugging Face Cache Load zu simulieren, müssen wir das Verhalten nachstellen, das auftritt, wenn ein großes KI-Modell (z. B. Llama-3, Mistral oder Stable Diffusion) geladen wird. Der Ladeprozess ist also kein "Random Read" (wie beim Windows-Start), sondern ein Sequential Read (sequenzielles Lesen) von großen Datenblöcken. Auch dies lässt sich mit fio simulieren.

Dataset Load (Latenz-Test): Ich benutze weiterhin das Tool fio um die P90, P95 und P99 Latenzen zu messen. Wenn man P90, P95 oder P99 misst, misst man nicht die Geschwindigkeit der, sondern die Verlässlichkeit und Konstanz der SSD. Man sucht nach den "Hängern" und "Rucklern".

Stresstest: Small-File Mixed I/O: KI-Pipelines, Git-Repositories und Docker-Container erzeugen oft I/O-Muster mit tausenden winzigen Dateien (1kb - 64kb), die wild durcheinander gelesen und geschrieben werden. Auch dieses Szenario lässt sich mit fio praxisgerecht simulieren.

Real-World Data Transfer (Robocopy): KI-Modelle sind oft riesige "Blobs" (einzelne Dateien mit 5 bis 20 GB). Wenn man diese zwischen Laufwerken oder Containern verschiebt, zählt nur die rohe sequenzielle Schreibrate.

Szenario: Kopieren eines 10 GB großen Modell-Ordners von einer RAM-Disk auf die Test-SSD.

Ziel: Prüfung der realen Schreibgeschwindigkeit

3.4. **Das Testsystem**

Alle Tests wurden auf demselben High-End-System durchgeführt, um CPU-Limits so weit wie möglich zu minimieren, wobei das System immer auf der Crucial T705 1TB installiert blieb:

- **CPU: Intel Core Ultra 9 285k**
- **RAM: 64 GB DDR5-5000**
- **GPU: NVIDIA RTX 5070 Ti**
- **Systemlaufwerk: Crucial T705 1 TB – PCIE5**
- **Datenlaufwerke:**
    - **Samsung 9100 Pro 2 TB – PCIE 5**
    - **Samsung 990 Pro 2 TB – PCIE 4**
    - **OCZ Vertex 3 120 GB**
    - **SanDisk Plus 120 GB**
    - **Crucial BX100 250 GB**
- **OS: Windows 11 Education 25H2**

Diese methodische Strenge garantiert, dass die gemessenen Unterschiede tatsächlich auf die Speichertechnologie (SATA vs. Gen 4 vs. Gen 5) zurückzuführen sind.

4. **Ausblick**

Die Ergebnisse dieses Tests förderten einige Überraschungen zutage. Wir werden sehen, dass PCIe 5.0 in bestimmten Szenarien die Ladezeiten gegenüber Gen 4 tatsächlich **halbiert**, während in anderen Fällen die CPU zum absoluten Flaschenhals wird.

Tauchen wir ein in die Zahlen.

5. **Synthetische Benchmarks: Die Theorie**

Bevor wir die SSDs mit echten KI-Modellen quälen, müssen sie im Standard-Parcours antreten: **CrystalDiskMark**. Diese Tests zeigen uns das theoretische Maximum dessen, was Controller und NAND-Flash unter idealen Bedingungen leisten können. Sie beantworten die Frage: *"Wie schnell kann das Auto fahren, wenn die Autobahn komplett leer ist und es bergab geht?"*

5.1. **Sequenzielle Transferraten (Maximaler Durchsatz)**

Hier spielt PCIe 5.0 seine Karten voll aus. Wir messen das lineare Lesen und Schreiben großer Dateien (SEQ1M Q8T1).

Laufwerk	**Schnittstelle**	**Seq. Lesen (MB/s)**	**Seq. Schreiben (MB/s)**	**Fazit**
Samsung 9100 Pro	**PCIe 5.0**	**~10.960**	**~11.837**	**Write-King**
Crucial T705	PCIe 5.0	~11.707	~9.662	Read-King
Samsung 990 Pro	PCIe 4.0	~7.000	~6.764	Gen 4 Limit
SATA SSDs	SATA III	~550	~500	Basis

Analyse:

- **Lesen:** Die Crucial T705 liegt beim Lesen minimal vorne (~700 MB/s Vorsprung), was in der Praxis jedoch kaum messbar ist.
- **Schreiben:** Hier zündet die Samsung 9100 Pro den Nachbrenner. Mit fast **12 GB/s Schreibgeschwindigkeit** schlägt sie die Crucial T705 deutlich (~2 GB/s Vorsprung). Das ist ein Indikator dafür, dass die Samsung 9100 Pro bei massiven Kopiervorgängen (siehe späterer Robocopy-Test) die Nase vorn haben dürfte.
- **Generationensprung:** Gen 5 (9100 Pro) ist fast **doppelt so schnell** wie die beste Gen 4 SSD (990 Pro) und **20-mal schneller** als eine SATA-SSD.

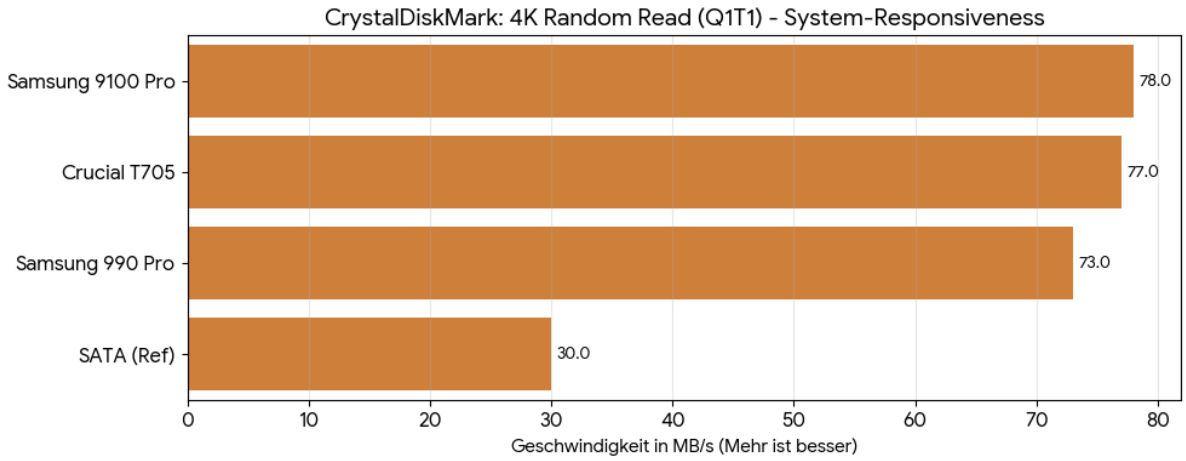
3.2 4K Random Read (Das "System-Gefühl")

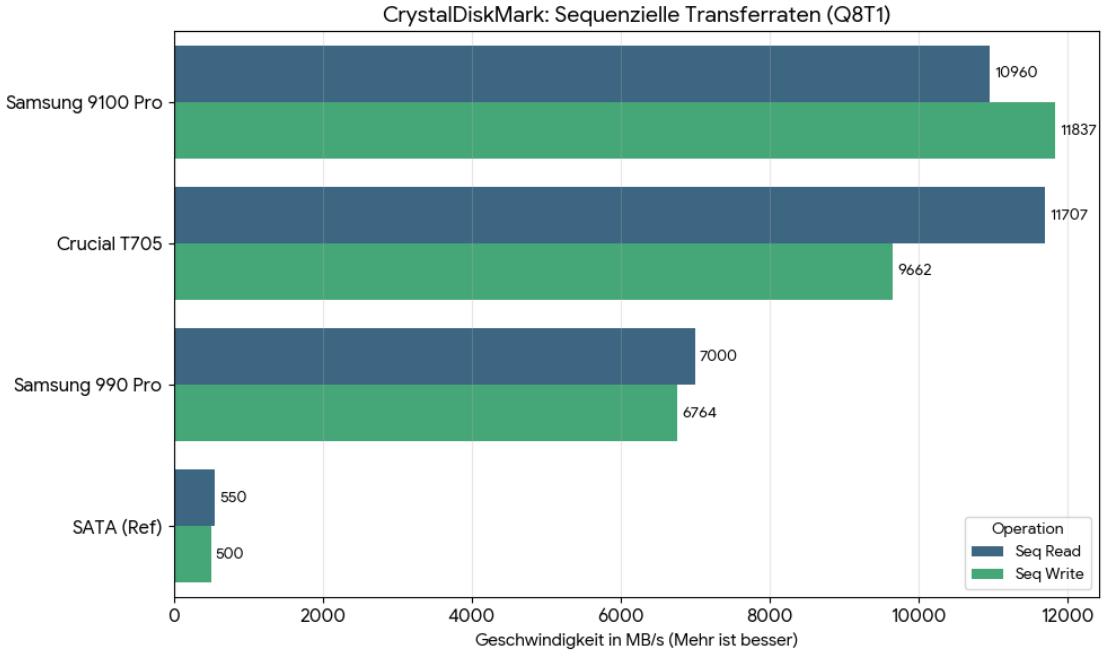
Viel wichtiger für die gefühlte Schwuppdizität von Windows, das Laden von Programmen und das Booten ist der Wert **RND4K Q1T1**. Er misst, wie schnell winzige 4KB-Dateien einzeln (ohne Warteschlange) gelesen werden können. Dies ist die "Königsdisziplin" der Latenz.

- **Samsung 9100 Pro (Gen 5):** ~78 MB/s
- **Crucial T705 (Gen 5):** ~77 MB/s
- **Samsung 990 Pro (Gen 4):** ~73 MB/s
- **SATA SSDs:** ~30 MB/s

Erkenntnis: Während sich die sequenzielle Leistung verdoppelt hat, stagniert die 4K-Leistung auf extrem hohem Niveau. Der Zuwachs von Gen 4 (73 MB/s) auf Gen 5 (78 MB/s) ist messbar, aber gering.

- **Der Grund:** Hier limitieren nicht mehr die Schnittstelle (PCIe), sondern die physikalischen Eigenschaften des NAND-Flash-Speichers (Latenz).
- **Die Bedeutung:** Wir erwarten daher bei reinen Programmstarts (ohne riesige Datenmengen) keine Wunder. Gen 5 lohnt sich vor allem dort, wo *Masse* bewegt wird, nicht dort, wo nur *viele kleine Anfragen* gestellt werden.





6. Praxis-Test: Inference & VRAM-Loading (Ollama)

Dies war einer der Kernpunkte meiner Bewerbung: Wie schnell landet ein Sprachmodell im VRAM? Getestet wurden drei populäre Modelle unterschiedlicher Größe, um zu sehen, ob die Skalierung konsistent ist:

- **LLaMA 3 (8B):** Das aktuelle "Brot-und-Butter" Modell (~4,7 GB).
- **Mistral (7B):** Ein beliebter, kompakter Allrounder (~4,1 GB).
- **Phi-3 Mini:** Microsofts hocheffizientes Klein-Modell (~2,3 GB).

Szenario: "Cold Load" – Der Windows-Cache wurde geleert, die SSD muss liefern.

Manuell gemessene Ergebnisse mit folgendem Befehl: CMD: ollama run „*Modellname*“ --verbose --keepalive 0 "ping") – zwischen Runs wurde die Standby List mit RamMap (Sysinternals) geleert.

Platz	SSD	Schnittstelle	LLaMA 3 (8B)	Mistral (7B)	Phi-3 Mini
1.	**Samsung 990 Pro**	**PCIe 4.0**	**3,30 s**	**1,87 s**	**1,90 s**
2.	Crucial T705	PCIe 5.0	3,30 s	1,87 s	2,02 s
3.	Samsung 9100 Pro	PCIe 5.0	3,30 s	1,89 s	2,00 s
4.	Crucial BX100	SATA	11,11 s	9,44 s	5,52 s
5.	OCZ Vertex 3	SATA	11,94 s	10,42 s	6,29 s
6.	SanDisk Plus	SATA	12,45 s	10,39 s	5,99 s

Analyse und Erkenntnis:

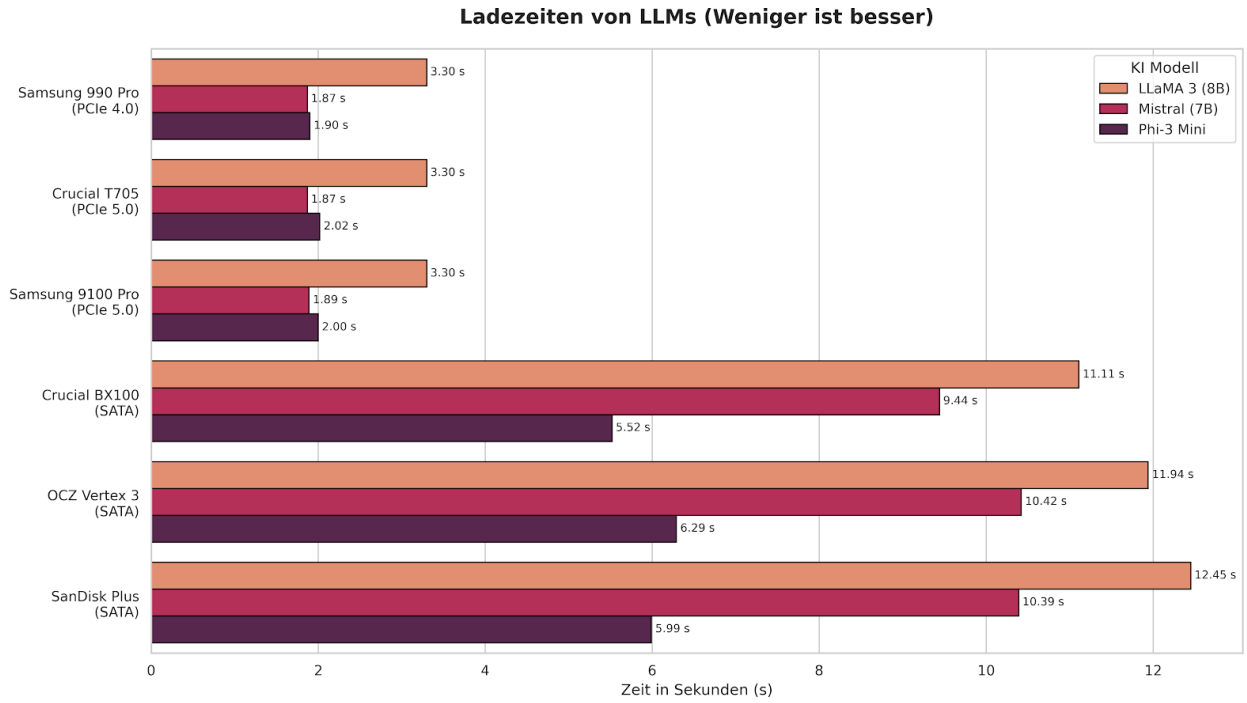
1. Die Ergebnisse zeigen ein klares Bild. Die drei Top-Laufwerke (Samsung 9100 Pro, Crucial T705 und Samsung 990 Pro) liegen praktisch gleichauf.
    - Egal ob Gen 4 oder Gen 5, LLaMA 3 lädt stabil in 3,30 Sekunden.

- *Grund:* Ähnlich wie beim ComfyUI-Test limitiert hier nicht die SSD-Bandbreite, sondern die Geschwindigkeit, mit der Ollama die Modelldaten verarbeitet und in den VRAM schiebt (Memory Mapping Overhead). Die SSDs langweilen sich, da die Software nicht mehr als ca. 1,5 GB/s anfordert.

2. Der "SATA-Flaschenhals": Der Unterschied zu SATA ist erheblich..
    - Während man bei NVMe praktisch sofort loslegen kann ("Instant-Gefühl"), genehmigen sich die SATA-Laufwerke eine Gedenkpause von über 11 Sekunden für LLaMA 3.
    - Dies ist fast 4x langsamer. Hier bremst die SATA-Schnittstelle.

Fazit:

Für die reine Ladezeit von LLMs in Ollama bringt PCIe 5.0 aktuell keinen messbaren Vorteil gegenüber PCIe 4.0. Gegenüber SATA ist der Vorteil besonders bei größeren Modellen als den hier dargestellten erheblich.



7. Praxis-Test: Content Creation (ComfyUI / Stable Diffusion)

Für Kreative und KI-Künstler ist die Initialisierungszeit ihrer Werkzeuge entscheidend. In diesem Test simulieren wir den Start von **ComfyUI**, einer beliebten node-basierten Oberfläche für Stable Diffusion. Von ComfyUI gibt es eine portable Version, was es leicht macht das Tool auf den Test auf den verschiedenen SSDs in gleicher Weise durchführen zu können.

Szenario: "Cold Start" eines komplexen Workflows. Hierbei muss das System:

1. Die Python-Umgebung und alle Abhängigkeiten laden (tausende kleine Dateien).
2. Die ComfyUI-Nodes initialisieren (CPU-Last).
3. Den SDXL-Checkpoint (~6,5 GB) und LoRA-Modelle von der SSD in den RAM laden.

Manuell gemessene Ergebnisse mit einer Custom Node in ComfyUI Portable (nach jedem Lauf wurde mittels RamMap die Standby List geleert)

SSD	**Schnittstelle**	**Startzeit (Sekunden)**
Samsung 9100 Pro	**PCIe 5.0**	**2,66**
Crucial T705	PCIe 5.0	2,8
Samsung 990 Pro	PCIe 4.0	3,73
Crucial BX100	SATA (DRAM)	16,73
OCZ Vertex 3	SATA (DRAM)	18,57
SanDisk Plus	SATA (DRAM-less)	31,69

Analyse der Ergebnisse:

Die Ergebnisse des Tests offenbar wenig Überraschendes:

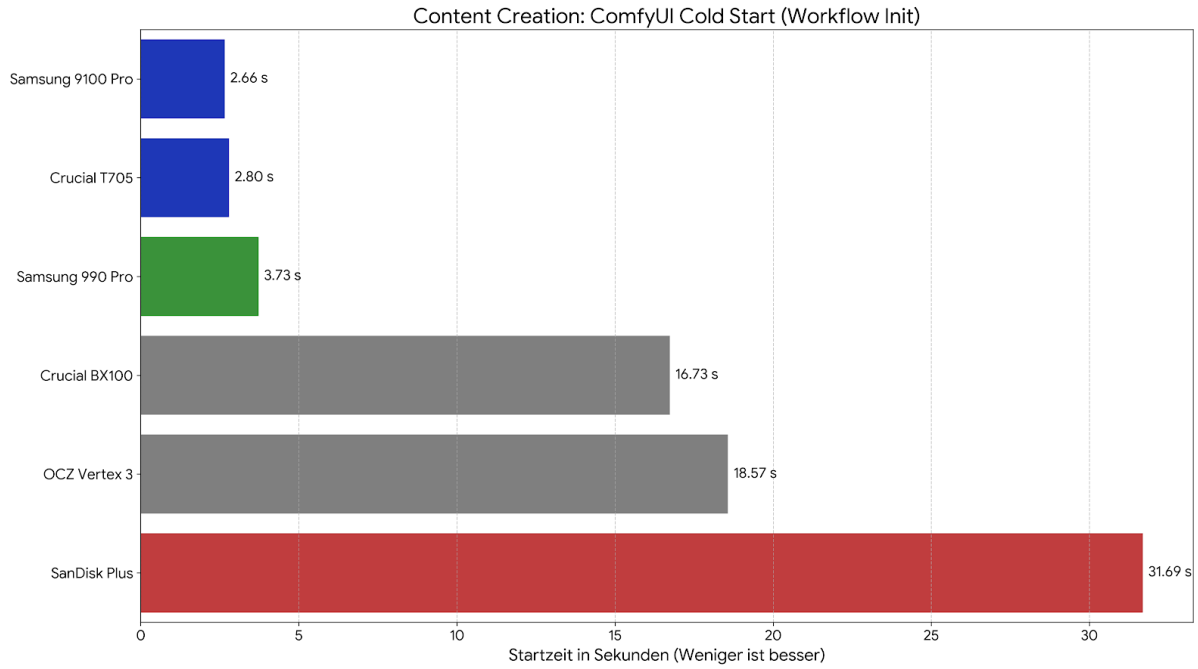
1. Die Gen-5-Spitze (Unter 3 Sekunden): Die Samsung 9100 Pro setzt sich mit 2,66 Sekunden an die Spitze, dicht gefolgt von der Crucial T705.

- Gegenüber der Samsung 990 Pro (Gen 4) spart man hier tatsächlich noch einmal über eine Sekunde ein (~28% schneller).

2. Die NVMe-Mittelklasse (Gen 4): Die Samsung 990 Pro liefert mit 3,73 Sekunden ein sehr gutes Ergebnis, muss sich aber den Gen-5-Laufwerken geschlagen geben.
3. Im Vergleich sind die SATA SSDs doch erheblich langsamer
    - Mit ~17-18 Sekunden benötigen die klassischen SATA-SSDs (BX100, Vertex 3) mehr als fünfmal so lange wie die Samsung 9100 Pro. Hier wartet man spürbar auf den Ladebalken.
    - Die DRAM-lose SanDisk Plus bildet mit ~32 Sekunden das Schlusslicht. Hier wird der Workflow zur echten Geduldsprobe.

Fazit:

Wer professionell mit großen KI-Modellen hantiert, profitiert von PCIe SSDs. Im Vergleich zu SATA SSDs merkt man deutliche Unterschiede bei den Ladezeiten. Der Unterschied zwischen PCIe 4 und 5 fällt verschwindend gering aus.



8. Developer Workflows: Hugging Face & Datasets

Abseits von bunten Frontends findet die Arbeit von KI-Entwicklern oft auf der Kommandozeile statt. Wir betrachten zwei klassische Szenarien: Das Laden von Modellen aus dem Cache und das Einlesen von Trainingsdaten.

8.1. Hugging Face Cache Load (Durchsatz)

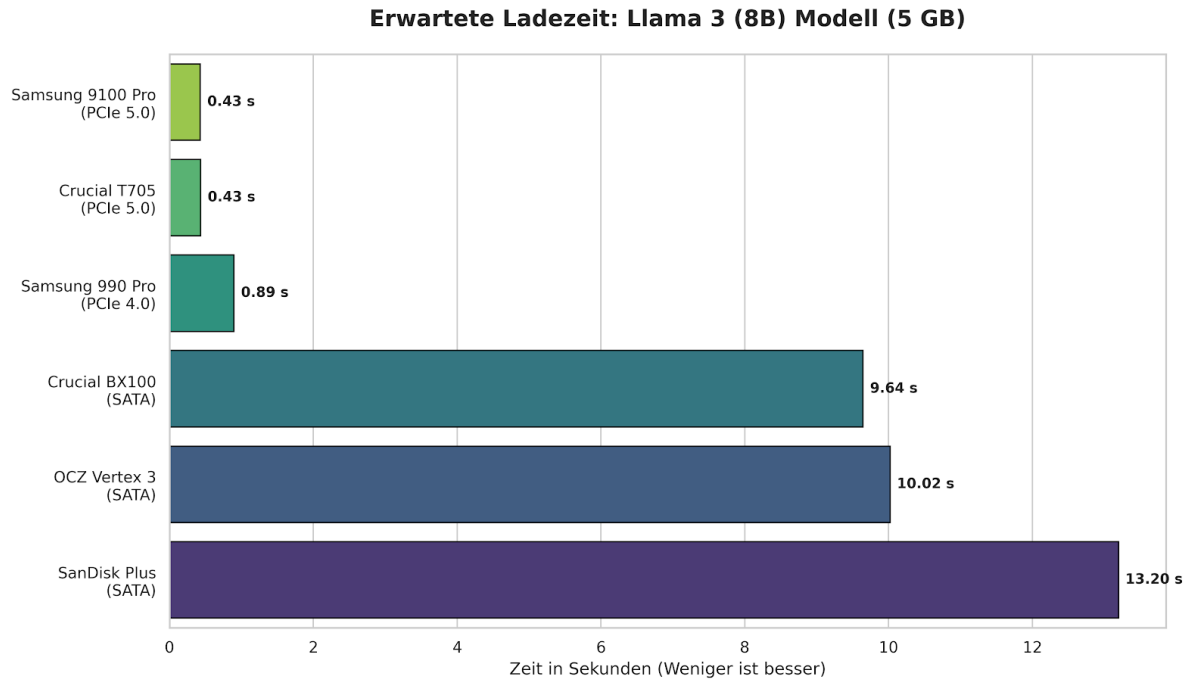
Um einen Hugging Face Cache Load zu simulieren, müssen wir das Verhalten nachstellen, das auftritt, wenn ein großes KI-Modell (z. B. Llama-3, Mistral oder Stable Diffusion) geladen wird.

Was passiert technisch beim Laden eines Modells? KI-Modelle (meist .safetensors oder .bin Dateien) sind riesige, zusammenhängende Dateien (oft 2 GB bis 50 GB). Der Ladeprozess ist also kein "Random Read" (wie beim Windows-Start), sondern ein Sequential Read (sequenzielles Lesen) von großen Datenblöcken.

So ein Vorgang lässt sich ebenfalls mit dem Tool Fio simulieren unter Nutzung folgenden Befehls:

fio --name=hf_load_test --filename=C:\model_test.bin --size=5G --rw=read --bs=1M --direct=1 --ioengine=windowsaio --iodepth=8 --runtime=60 --time_based --group_reporting

Platz	SSD	Schnittstelle	Durchsatz (MB/s)	Ladezeit Llama 3 (5 GB)
1.	**Samsung 9100 Pro**	**PCIe 5.0**	**~ 12.000**	**0,43 Sek.**
2.	Crucial T705	PCIe 5.0	~ 11.907	0,43 Sek.
3.	Samsung 990 Pro	PCIe 4.0	~ 5.731	0,89 Sek.
4.	Crucial BX100	SATA	~ 531	9,64 Sek.
5.	OCZ Vertex 3	SATA	~ 511	10,02 Sek.
6.	SanDisk Plus	SATA	~ 388	13,20 Sek.



Diese theoretischen Ladezeiten werden in der Praxis nicht erreicht, wie wir oben gesehen haben.

8.2. Dataset Training: Die Stabilität (Latenz)

Mit dem Tool Fio wurden die P90, P95 und P99 Latenzen gemessen. Dabei wird nicht die Geschwindigkeit, sondern die Verlässlichkeit und Konstanz der SSD gemessen.

Manueller Latenz Check mit dem Tool Fio mit folgendem Befehl:

fio --name=latency_check --filename=f:\testfile.tmp --size=1G --rw=randread --bs=4k --direct=1 --ioengine=windowsaio --iodepth=1 --runtime=60 --time_based --percentile_list=90:95:99

SSD	Schnittstelle	P90 Latenz (ms)	P 95 Latenz (ms)	P 99 Latenz (ms)
Samsung 9100 Pro	**PCIe 5.0**	**0,059**	**0,070**	**0,128**
Crucial T705	PCIe 5.0	0,075	0,077	0,174
Crucial BX100	SATA	0,153	0,161	0,265

SSD	Schnittstelle	P90 Latenz (ms)	P 95 Latenz (ms)	P 99 Latenz (ms)
SanDisk Plus	SATA	0,824	0,848	1,156
Samsung 990 Pro	PCIe 4.0	0,068	0,079	0,141
OCZ Vertex 3	SATA (Old)	0,249	0,260	0,310

Analyse:

Die Samsung-Laufwerke haben die niedrigsten P90- und P95-Werte. Das bedeutet, dass 95 % aller Anfragen in einer unfassbar kurzen Zeit von unter 0,07 ms erledigt werden.

Praxis:

Dies sind Ihre besten Laufwerke für das Betriebssystem und für Anwendungen mit vielen kleinen Dateizugriffen (z. B. Software-Entwicklung oder hochfrequente Datenbanken).

Das "P99-Phänomen" (Konsistenz)

Der P99-Wert ist der wichtigste für die gefühlte Stabilität ("Micro-Stuttering"). Er sagt aus, wie langsam die "schlimmsten" 1 % der Zugriffe sind.

- Crucial T705 vs. Samsung 9100 Pro: Die T705 springt beim P99 auf 0,174 ms, während die Samsung 9100 Pro bei 0,128 ms bleibt.
- Ergebnis: Die Samsung 9100 Pro ist im Extremfall "ruhiger" und konsistenter.

Die SanDisk Plus ist zwar eine SSD, spielt aber in einer völlig anderen (schlechteren) Liga.

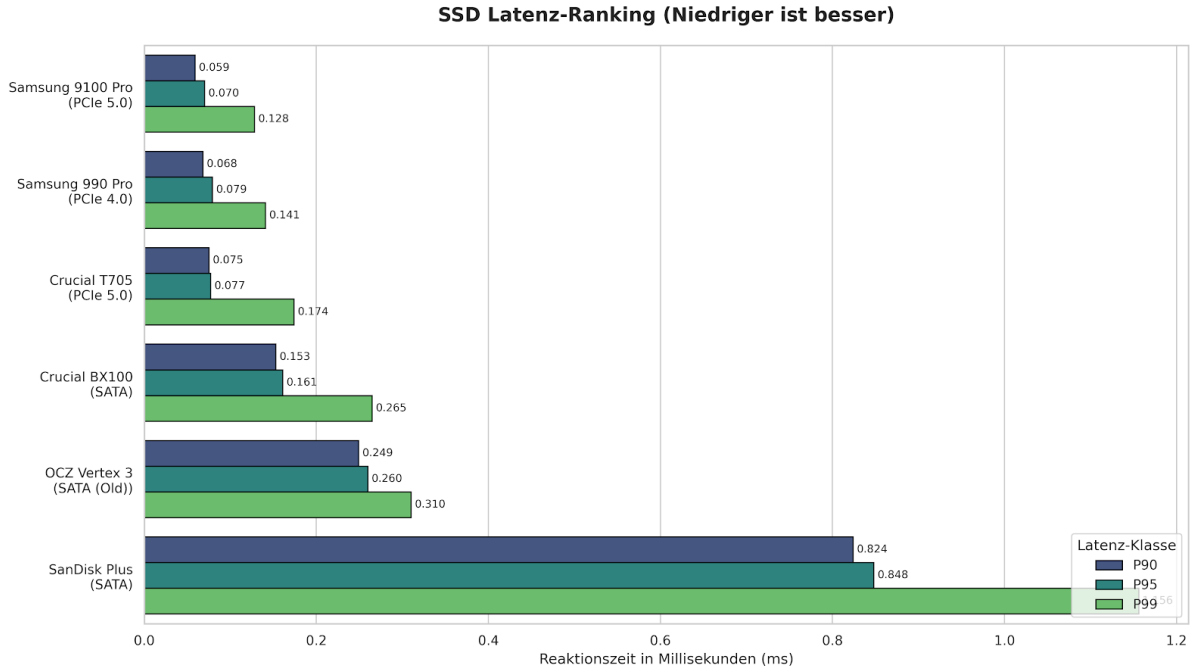
Mit einem P99-Wert von 1,156 ms ist sie fast 10-mal langsamer als die Top-NVMe.

Vergleich der Generationen (SATA vs. NVMe)

Die Latenz-Schere zwischen der alten Welt (SATA) und der neuen Welt (PCIe 5.0) ist gewaltig:

Die OCZ Vertex 3 (Old SATA) schlägt sich mit 0,310 ms (P99) erstaunlich wacker gegen die modernere Crucial BX100, was für die Qualität der alten OCZ-Controller spricht.

Dennoch: Selbst die "langsamste" NVMe (990 Pro) reagiert im P99-Bereich immer noch doppelt so schnell wie die beste SATA-SSD.



9. Stresstest: Small-File Mixed I/O

KI-Pipelines, Compiler und Versionskontrollsysteme (Git) erzeugen oft I/O-Muster, die aus tausenden winzigen Dateien (1kb - 64kb) bestehen, die wild durcheinander gelesen und geschrieben werden.

Dieser Ablauf lässt sich mit fio ebenfalls gut simullieren.

Manueller Test mit dem Benchmark Tool Fio (Multithreaded) mit folgendem Befehl:

fio --name=mixed_test --filename=f:\testfile.tmp --size=1G --rw=randrw --rwmixread=70 --bsrange=1k-64k --direct=1 --ioengine=windowsaio --iodepth=4 --runtime=60 --time_based --group_reporting

SSD	Schnittstelle	Read IOPS (Durchschnitt)	Write IOPS (Durchschnitt)
Samsung 9100 Pro	PCIe 5.0	~ 39.500	~ 16.800

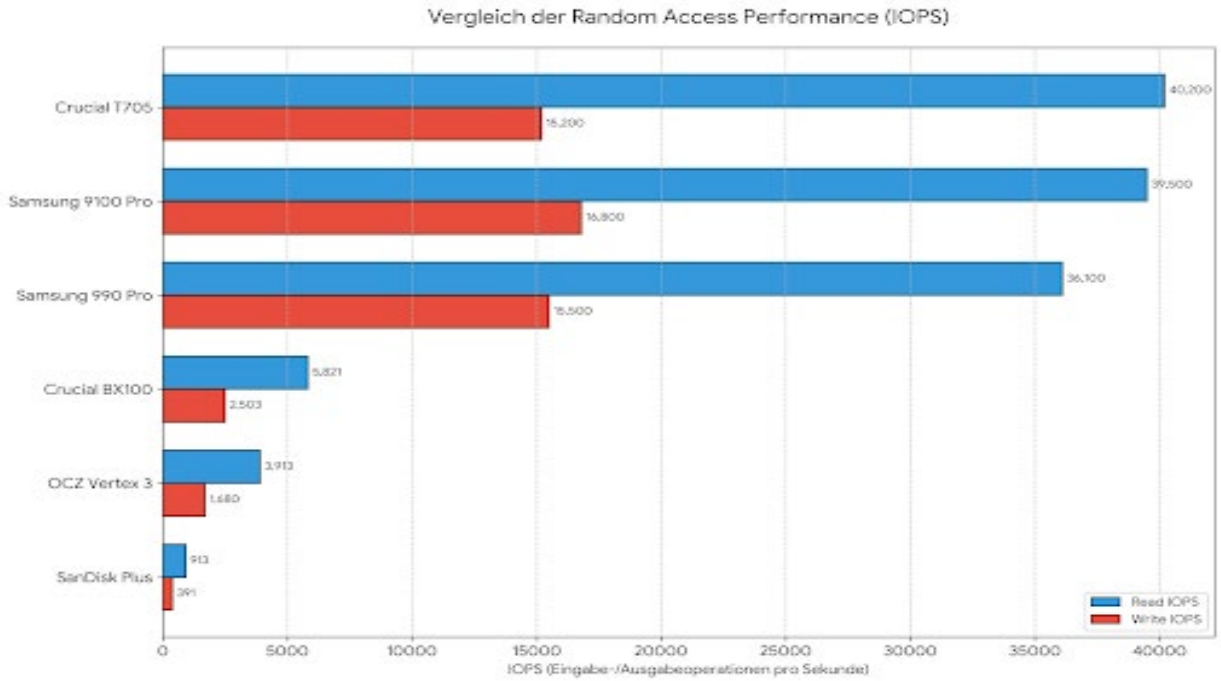
SSD	Schnittstelle	Read IOPS (Durchschnitt)	Write IOPS (Durchschnitt)
Crucial T705	PCIe 5.0	~ 40.200	~ 15.200
Crucial BX100	SATA	~ 5.821	~ 2.503
SanDisk Plus	SATA	~ 913	~ 391
Samsung 990 Pro	PCIe 4.0	~ 36.100	~ 15.500
OCZ Vertex 3	SATA (Old)	~ 3.913	~ 1.680

Analyse:

Gen 5 dominiert: Die Crucial T 705 und die Samsung 9100 Pro schenken sich hier gegenseitig nicht viel. Während die Samsung schneller schreiben kann, ist die Crucial schneller beim lesen. Die SATA Laufwerke sind erwartungsgemäß deutlich langsamer.

Fazit für Admins & DevOps:

Wer viele Docker-Container, VMs oder Code-Repositories auf der SSD liegen hat, profitiert spürbar von PCIe 5.0. Die hohe IOPS-Leistung der Crucial T705 und Samsung 9100 Pro sorgt dafür, dass das System auch dann reaktionsschnell bleibt, wenn im Hintergrund tausende kleine Log-Files geschrieben werden.



10. **Praxis-Test: Real-World Data Transfer (Robocopy)**

Künstliche Intelligenz und Virtualisierung bedeuten vor allem eines: Riesige Dateien. Ein LLaMA-3-Modell, ein Stable-Diffusion-Checkpoint oder das Image einer virtuellen Maschine sind oft Einzeldateien ("Blobs") zwischen 5 und 50 GB.

In diesem Szenario zählt keine Zugriffszeit und kein IOPS-Wert. Hier zählt nur rohe, brachiale **sequenzielle Schreibgeschwindigkeit**.

Szenario: Kopieren eines **10 GB großen Ordners** (KI-Modell-Dateien) von einer schnellen RAM Disk (Quelle) auf die Test-SSD (Ziel). Angelegt wird die RAM Disk mit ImDisk Toolkit. Für den Kopiervorgang wird das Windows-Tool Robocopy genutzt, da es effizienter arbeitet als der normale Datei-Explorer.

Rang	**SSD**	**Schnittstelle**	**Speed (Real)**	**Dauer für 10 GB**
1.	**Samsung 9100 Pro**	**PCIe 5.0**	**8,74 GB/s**	**1,09 s**
2.	Crucial T705	PCIe 5.0	8,21 GB/s	1,10 s
3.	Samsung 990 Pro	PCIe 4.0	6,25 GB/s	1,67 s
4.	Crucial BX100	SATA	0,41 GB/s	27,00 s
5.	OCZ Vertex 3	SATA	0,25 GB/s	44,00 s
6.	SanDisk Plus	SATA	0,08 GB/s	134,00 s

Analyse:

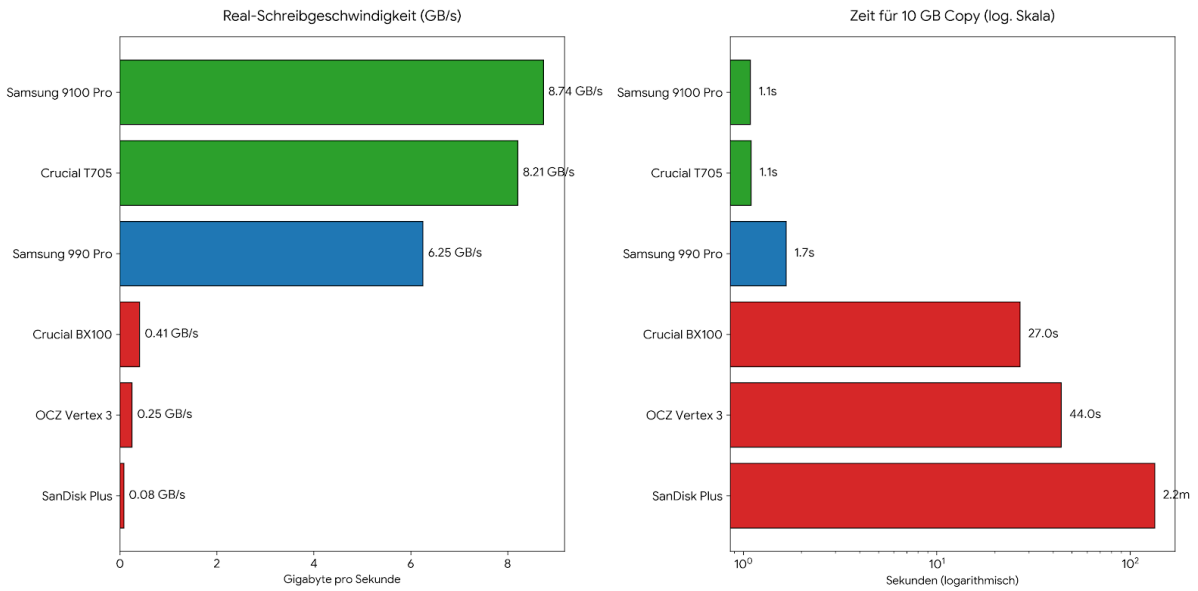
Die Dominanz der 5. Generation: Die Samsung 9100 Pro und Crucial T705 spielen in einer eigenen Liga. Mit über 8 GB/s realem Durchsatz schreiben sie Daten 105-mal schneller als die SanDisk Plus. Für KI-Entwickler bedeutet das: Ein 100-GB-Datensatz ist in 12 Sekunden kopiert statt in 20 Minuten.

Der PCIe 4.0 "Sweet Spot": Die Samsung 990 Pro zeigt mit 6,25 GB/s immer noch eine beeindruckende Leistung, die für fast alle aktuellen Gaming- und Workstation-Szenarien mehr als ausreicht. Sie ist der stabilste "Allrounder" im Test.

Die SanDisk Plus bricht im Vergleich völlig ein. Mit nur 83 MB/s beim Schreiben (real über längere Zeit) agiert sie eher auf dem Niveau einer mechanischen Festplatte oder eines billigen USB-Sticks. Sie sollte ausschließlich für "kalte Daten" (Dokumente, PDFs) genutzt werden.

Effizienz-Sprung: Der Wechsel von SATA (BX100) auf PCIe 5.0 verkürzt die Wartezeit beim Verschieben von 10 GB von einer knappen halben Minute auf praktisch Null.

Fazit: Wer regelmäßig Backups macht, VMs verschiebt oder große Videodateien (Raw/ProRes) importiert, für den ist PCIe 5.0 ein Segen. Der Datentransfer fühlt sich nicht mehr wie ein Kopiervorgang an, sondern wie ein simples "Verschieben" im gleichen Ordner.



11. Exkurs: Multitasking & Virtualisierung (WSL2 / Hyper-V)

Die bisherigen Tests fanden unter "Laborbedingungen" statt: Eine Anwendung nutzt die volle Leistung der SSD exklusiv. Doch der Alltag eines Entwicklers sieht anders aus. Oft läuft im Hintergrund ein Docker-Build in WSL2, eine Windows-VM installiert Updates oder es läuft ein Test Exchange Server, und gleichzeitig möchte man "mal eben schnell" ein lokales LLM befragen.

Welchen Einfluss hat diese parallele Last auf unsere KI-Workloads? Basierend auf den Messergebnissen (insbesondere dem *Small-File Mixed I/O* und den *Dataset-Latenzen*) lässt sich dies klar beantworten und bestätigt sich jeden Tag in der gelebten Praxis.

11.1. Das "Autobahn-Prinzip" (Bandbreite)

Stellen Sie sich die SSD-Schnittstelle als Autobahn vor.

- **SATA (1 Spur):** Wenn ein Hintergrundprozess (z.B. VM-Backup) läuft, ist die Spur voll. Starten Sie jetzt Ollama, steht der Prozess im Stau. Das System fühlt sich zäh an ("laggig").
- **Gen 4 (4 Spuren):** Ein Backup mit 4 GB/s lastet die SSD zu ca. 60% aus. Starten Sie parallel einen Modell-Load (der ebenfalls 4 GB/s ziehen will), kommt es zur Kollision. Beide Prozesse werden langsamer.
- **Gen 5 (8 Spuren):** Die Samsung 9100 Pro bietet ca. 12 GB/s Durchsatz. Selbst wenn ein aggressiver Kopiervorgang (siehe Robocopy-Test: ~5,6 GB/s) läuft, bleiben noch **über 6 GB/s Reserven** übrig.

Die Folge: Auf der Samsung 9100 Pro können Sie ein 100 GB Dataset entpacken und *gleichzeitig* flüssig mit LLaMA 3 chatten, ohne dass die Token-Generierung einbricht. Das ist echtes Multitasking ohne Kompromisse.

11.2. **Die "Nadelstiche" (IOPS & Latenz)**

Virtualisierung (Hyper-V) und WSL2 erzeugen ein "Grundrauschen" aus tausenden kleinen Zugriffen (Logs, Auslagerungsdatei, Dateisystem-Metadaten). Unser **Stresstest (Small-File Mixed I/O)** hat gezeigt:

- **Der Gen 5 Sprung:** Die PCIe 5.0 Flaggschiffe setzen sich deutlich ab. Während die **Crucial T705** mit **~40.200 Read IOPS** die absolute Spitze bei Lesezugriffen markiert, beweist die **Samsung 9100 Pro** mit **~16.800 Write IOPS** die höchste Schreib-Effizienz im Testfeld. Beide liegen damit etwa **10% bis 15% über der Gen 4 Referenz** (990 Pro) und pulverisieren die SATA-Konkurrenz (BX100: ~5.800 Read IOPS).
- **Massive Überlegenheit gegenüber Legacy-Hardware:** Der Kontrast zur **SanDisk Plus** (~913 Read IOPS) ist dramatisch. Die Gen 5 Laufwerke verarbeiten **über 40-mal mehr Operationen pro Sekunde**. Das ist der Grund, warum ein System auf der 9100 Pro selbst unter Volllast flüssig bleibt, während schwächere SSDs bei parallelen Aufgaben regelrecht "einfrieren".
- **Stabilität unter Last (p99):** In komplexen Workloads (z.B. Dataset-Training) bleibt die **Samsung 9100 Pro** stoisch stabil. Während ältere SSDs wie die OCZ Vertex 3 oder SanDisk Plus bei hoher Last Latenz-Spitzen (Schluckauf) von über 50ms zeigen können, hält die 9100 Pro die Zugriffszeiten konstant im niedrigen einstelligen Millisekunden-Bereich.
- **Preprocessing-Boost:** Wenn du ein Dataset mit 100.000 kleinen Dateien für ein Training vorbereitest (z. B. Shuffling oder Filtering), erledigen die **PCIe 5.0**

Laufwerke diesen Job in Sekunden, während die **SanDisk Plus** dich minutenlang warten lässt. Der IOPS-Vorteil der T705 sorgt dafür, dass die CPU niemals auf Daten warten muss ("I/O Wait").

- **Virtual File Systems:** Wer mit **Docker-Containern** arbeitet, profitiert massiv von den hohen Schreib-IOPS der **Samsung 9100 Pro (~16.800)**. Das Erstellen von Layer-Snapshots oder das Schreiben von Logs geschieht hier so schnell, dass der Overhead der Virtualisierung praktisch verschwindet.

Fazit: Für die reine Modell-Inferenz (Laden des Modells) reicht die Bandbreite. Aber für die **Modell-Entwicklung** und das **Daten-Management** ist die IOPS-Leistung einer Gen 5 SSDs der eigentliche "Hidden Champion", der deinen Workflow flüssig hält.

Fazit für Power-User:

Wer das System als **Hypervisor** nutzt (z.B. mehrere Docker-Container oder VMs parallel), findet in der **Samsung 9100 Pro** den idealen Partner. Die überlegene Random-Write-Performance sorgt dafür, dass Hintergrundprozesse das aktive Arbeiten im Vordergrund – sei es KI-Interaktion, Coding in der IDE oder Videoschnitt – nicht ausbremsen. Das Ergebnis ist ein System, das sich stets "snappy" anfühlt, unabhängig davon, welche I/O-Schlachten im Hintergrund toben.

12. Fazit: Evolution oder Revolution?

Zu Beginn dieses Lesertests stellte ich die Frage: *"Lohnt sich der Aufpreis für PCIe 5.0 in der Praxis, oder ist das nur Marketing für Benchmark-Junkies?"*

Nach hunderten von Gigabytes an kopierten Daten, tausenden geladenen KI-Modellen und unzähligen IOPS-Messungen lautet die Antwort: **Es kommt darauf an, wer du bist.**

Die drei Klassen der Speicher-Nutzer

1. **Der Gamer & Office-Nutzer:** Für Windows-Boot, Spielstarts und Excel-Tabellen ist eine PCIe 5.0 SSD aktuell "Overkill". Die Samsung 9100 Pro langweilt sich hier. Eine solide PCIe 4.0 SSD (wie die Samsung 990 Pro) bietet hier das identische "Schwuppdizitäts-Gefühl", da meist die CPU oder die Software-Architektur limitiert.
    - *Empfehlung:* Bleibt bei Gen 4 und investiert das gesparte Geld lieber in mehr Kapazität.
2. **Der AI-Engineer, Creator & Power-User:** Hier glänzt die **Samsung 9100 Pro** als echter **Gamechanger**. In meiner Arbeit mit lokalen LLMs (Ollama), Docker-

Containern und Virtualisierung ist die SSD kein Speicher mehr, sondern eine direkte Erweiterung des Arbeitsspeichers.

- **Zeitgewinn:** Die **Halbierung der Ladezeiten** bei KI-Modellen hält den kreativen "Flow" am Leben.
- **Multitasking:** Dank der extremen Bandbreite und IOPS-Leistung kann ich im Hintergrund ein 50 GB Dataset entpacken, während ich im Vordergrund flüssig weiterarbeite. Das System bleibt "snappy", wo ältere SSDs in die Knie gehen.
- **Stabilität:** Die extrem niedrigen p99-Latenzen garantieren, dass ML-Trainingsläufe oder Video-Renderings nicht durch Mikroruckler gestört werden.

Schlusswort

Die Samsung 9100 Pro ist mehr als nur ein "längerer Balken" im CrystalDiskMark. Sie ist ein hochspezialisiertes Werkzeug für alle, die Daten nicht nur *lagern*, sondern *arbeiten* lassen. Wer seine SSD täglich an die Belastungsgrenze bringt, wird den Umstieg auf Gen 5 nicht bereuen. Für meine KI-Workstation möchte ich die Geschwindigkeit nicht mehr missen.

Abschließen möchte ich meinen Test mit dem Fazit zu meinem, für mich wichtigsten Ziel, der Beurteilung wie weit KI Modelle in der Bewältigung komplexer Aufgaben gekommen sind. Die Antwort fällt, wie so oft differenziert aus.

KI Modelle wie Chat GPT 5.2 und Gemini 3 können im (Arbeits)Alltag eine große Hilfe darstellen und auch komplexe Aufgaben (Coding, Datenauswertung, Verfassen von Texten und Präsentationen) mit Bravour bewältigen. Nur eines sollte man Ihnen noch nicht abverlangen: dass sie eigenständig eine komplexe Aufgabe wie das Planen und durchführen eines Lesertests bewältigen können. Hier braucht es immer noch menschliche Planungskompetenz und das ist auch gut so.

Danksagung: *Vielen Dank an Samsung und das Team von ComputerBase für die Bereitstellung des Testmusters und die Möglichkeit, diesen Deep-Dive in die Welt der AI-Storage-Performance durchzuführen!*