

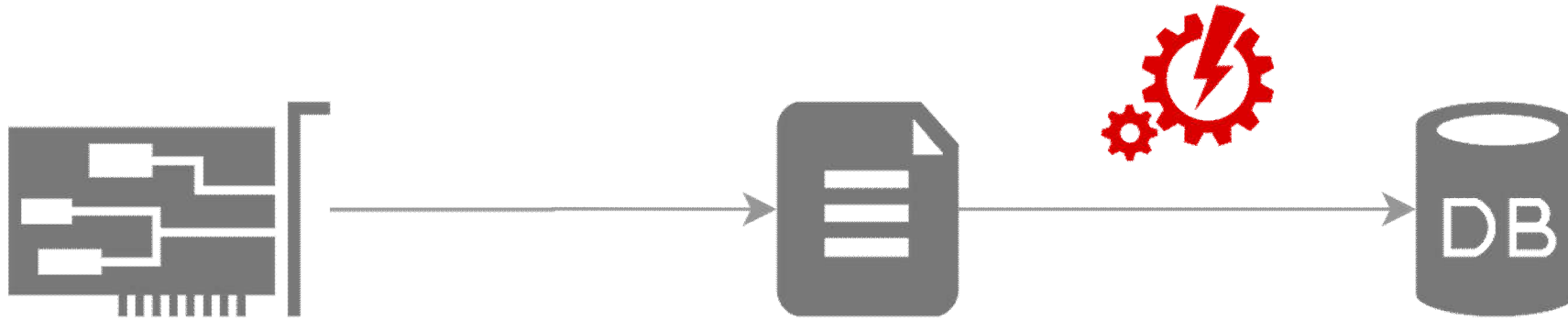
Fast CSV Loading Using GPUs and RDMA for In-Memory Data Processing

BTW 2021

Alexander Kumaigorodski, Clemens Lutz, Volker Markl

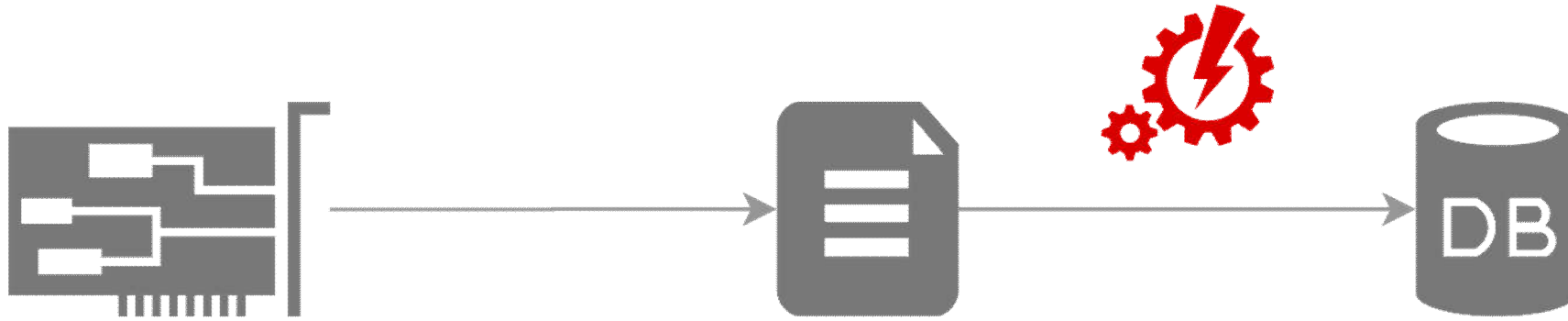
TU Berlin, DFKI

Goal



NVLink, RDMA, GPUDirect, ...
IO with 100+ Gbit/s

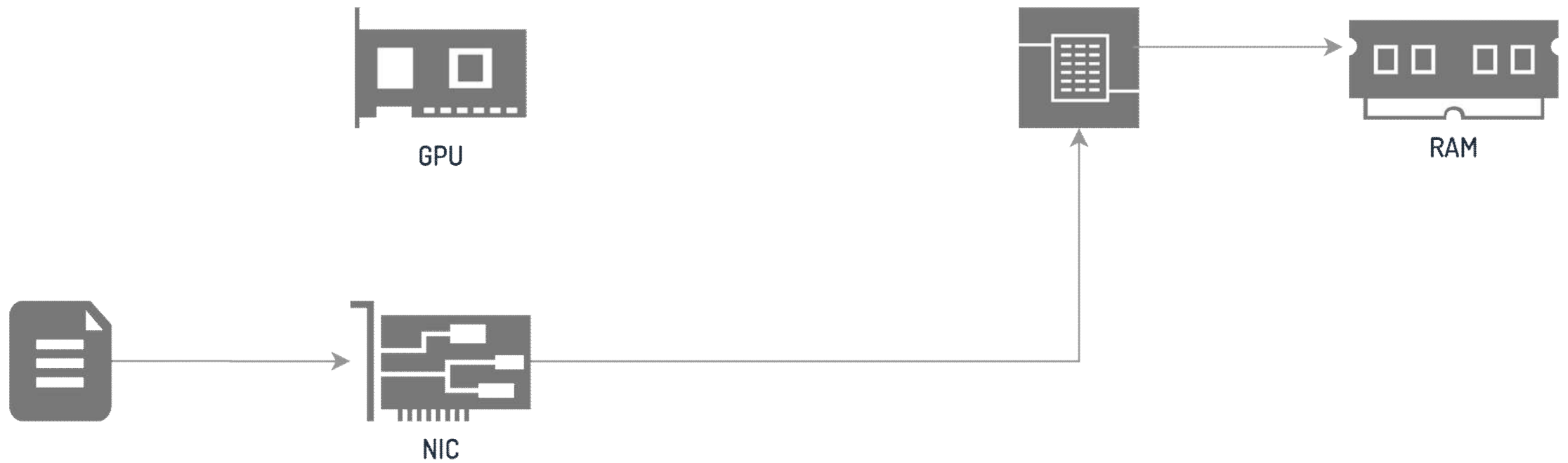
Goal



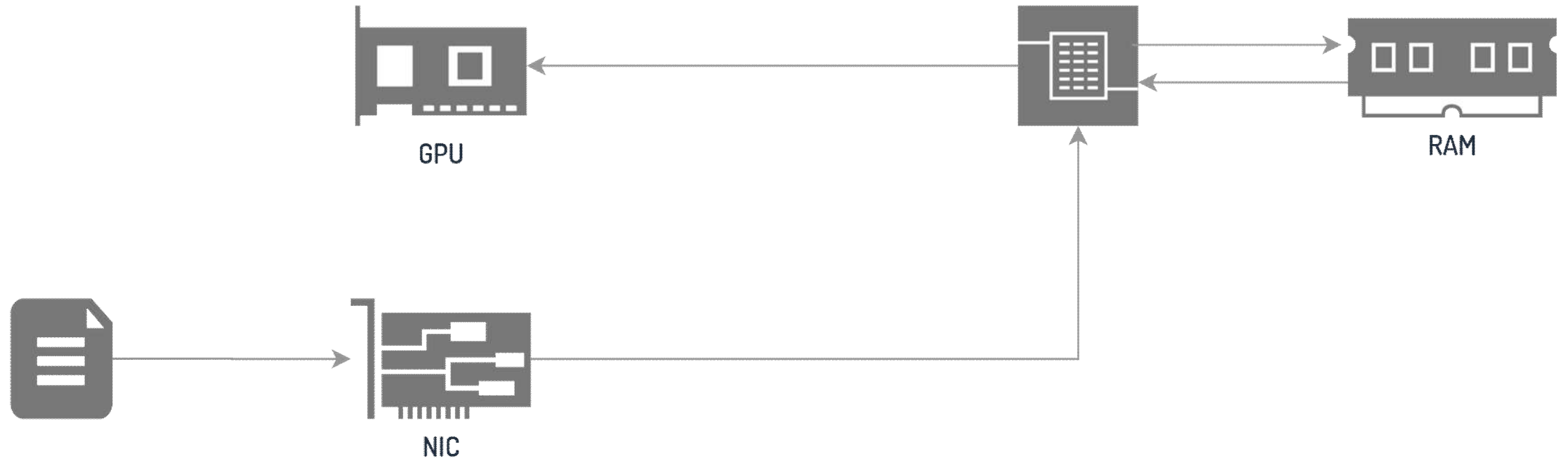
Efficiently load and parse
CSV files end-to-end

Goal

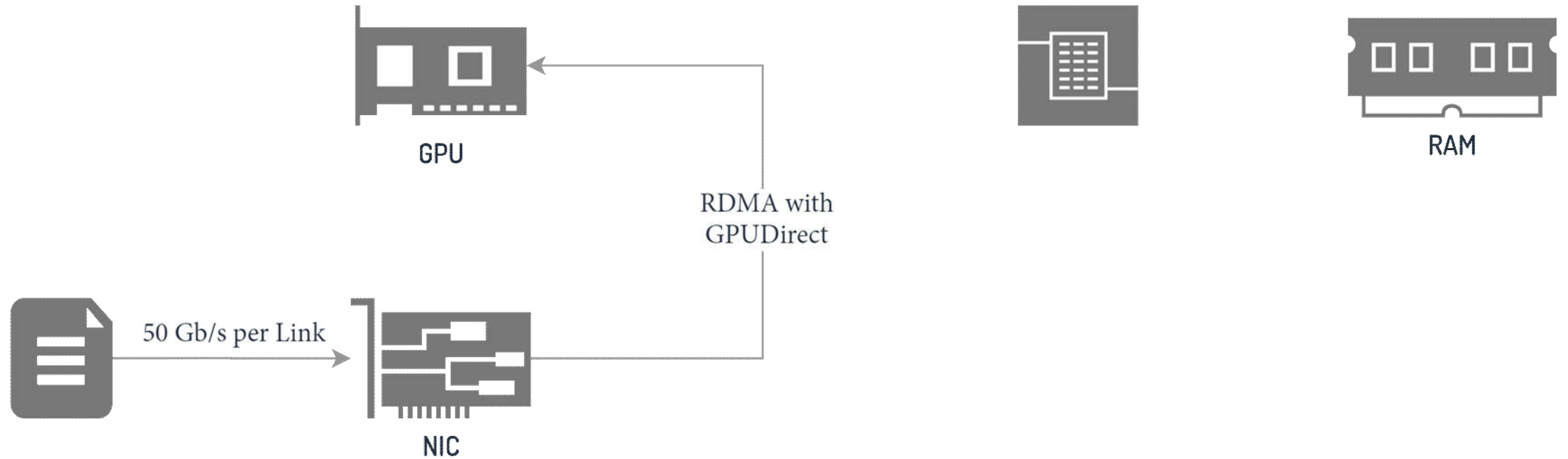
Goal



Goal



Goal



Contributions

Contributions

- Parsing on GPUs

Contributions

- Parsing on GPUs
- Streamed loading with GPUDirect RDMA

Contributions

- Parsing on GPUs
- Streamed loading with GPUDirect RDMA
- Evaluate fast interconnects for end-to-end streamed loading

Parsing on GPUs

Background: Warp Divergence

- No branch prediction on GPUs
- Threads execute in groups of 32, a *warp*
- Warp execute same instruction during cycle
- `if(cond) {...} else {...}` can cut performance in half

Problem: CSV Parsing

Problem: CSV Parsing

```
ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 47007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 126150745\n6, China, CN, 1427647786\n7, United States of America, US, 328239523\n8, Canada, CA, 37894799\n...
```

Parsing on GPUs

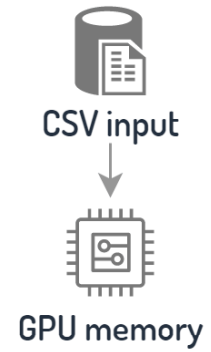
- Parsers typically have complex control flow (warp divergence)

Parsing on GPUs

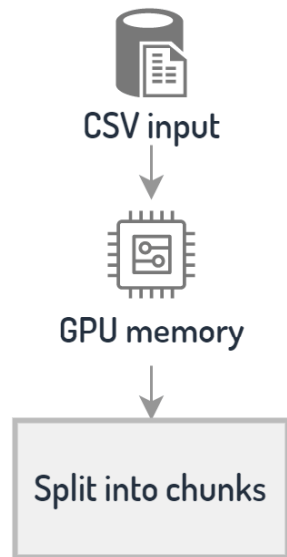
- Parsers typically have complex control flow (warp divergence)
- Exploring new trade-off:
Simplify control flow at expense of additional data passes (700+ GB/s)

Conceptual Overview

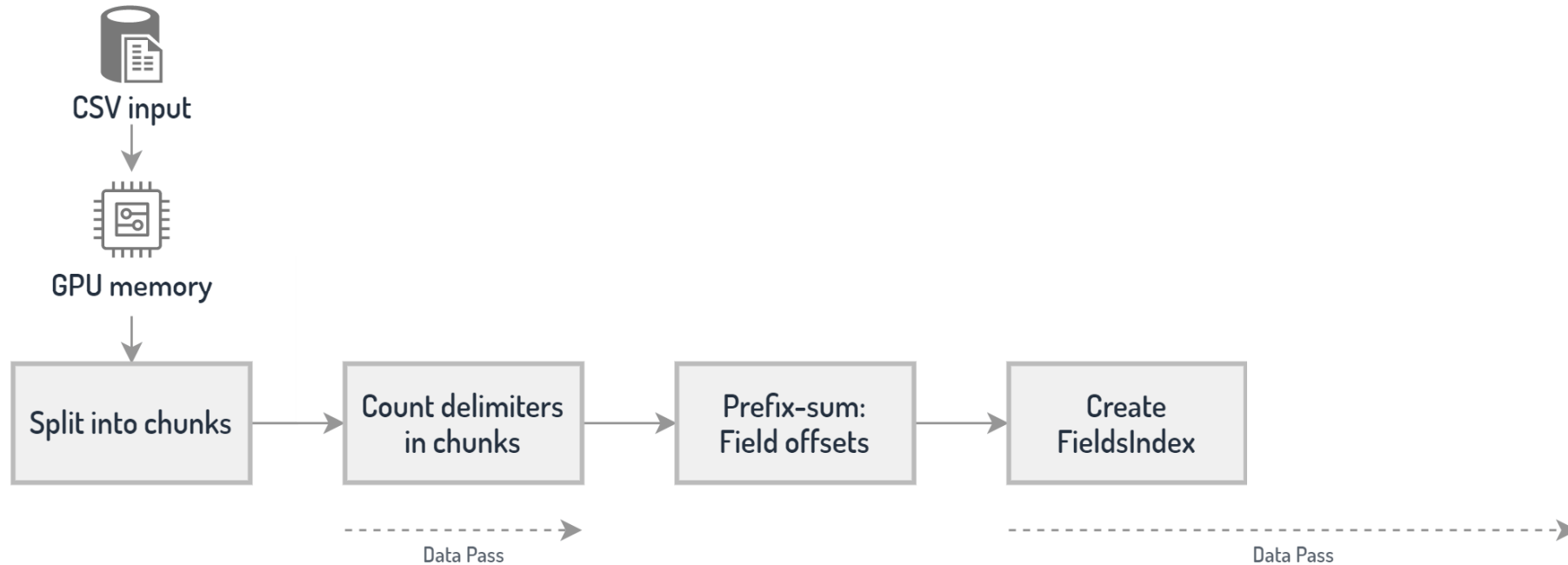
Conceptual Overview



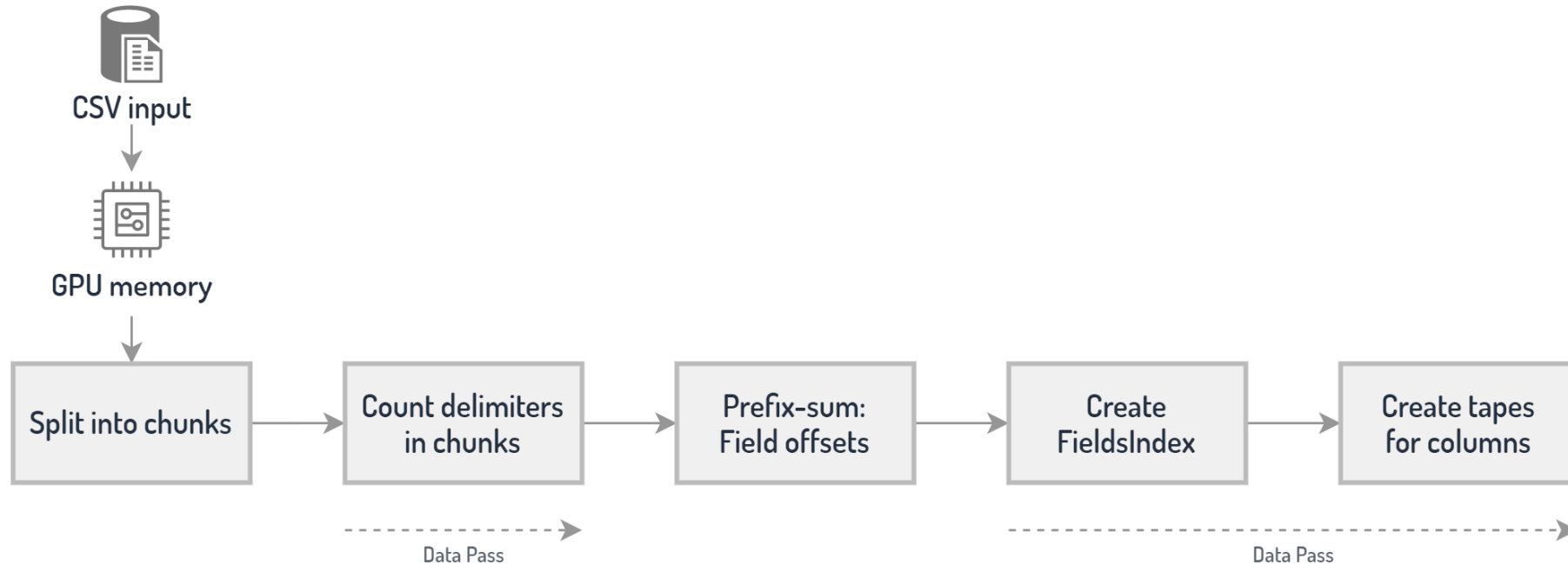
Conceptual Overview



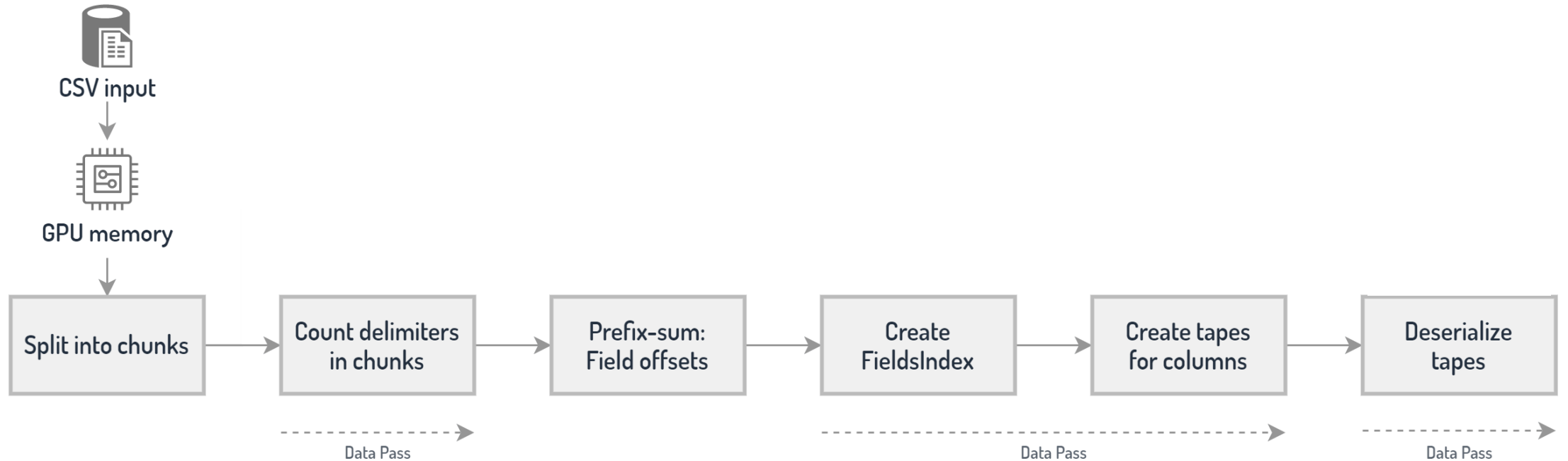
Conceptual Overview



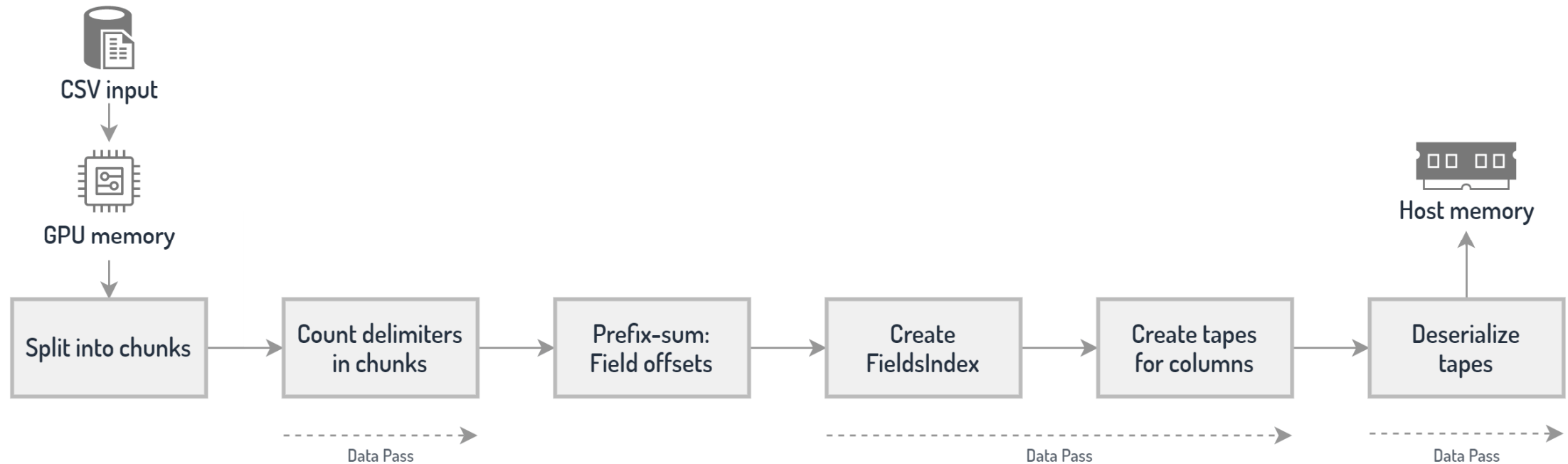
Conceptual Overview



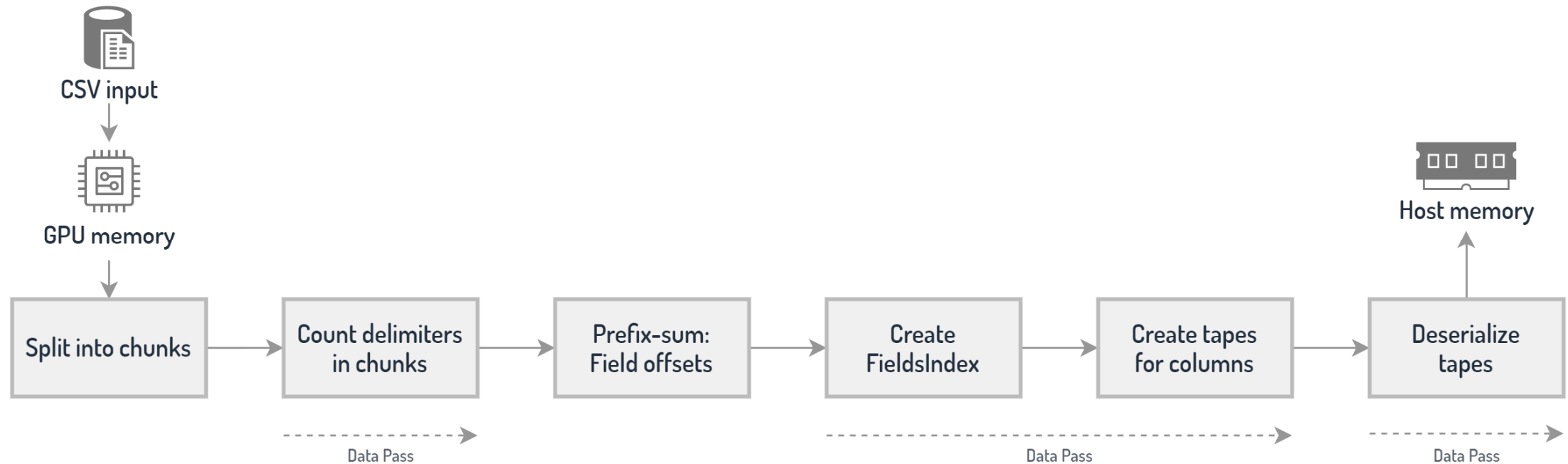
Conceptual Overview



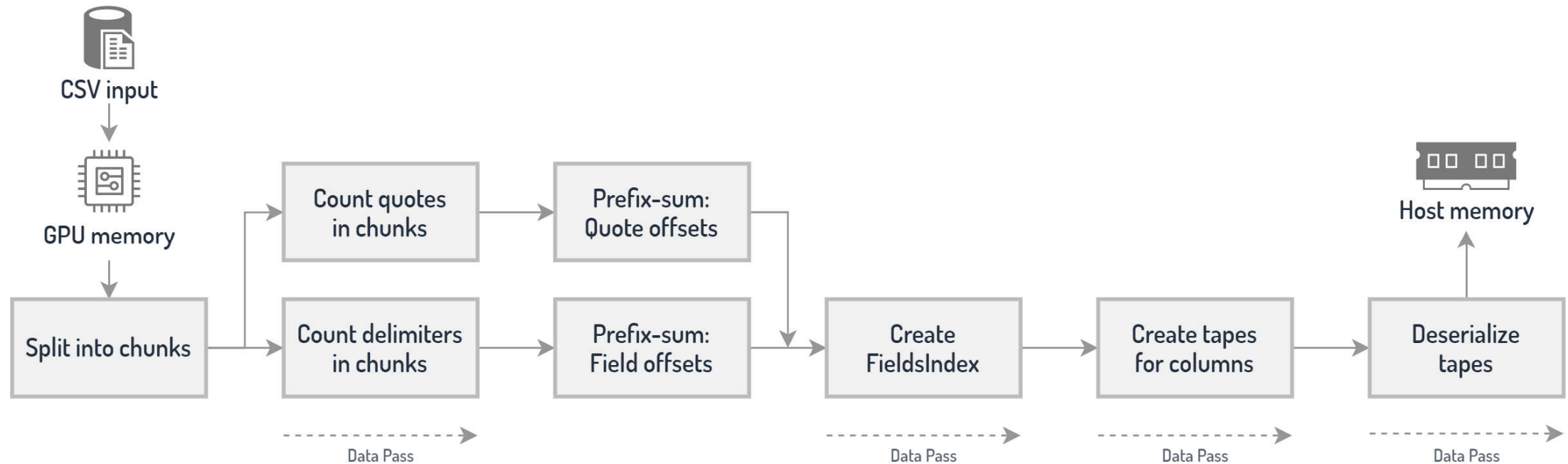
Conceptual Overview



Conceptual Overview: Fast Mode

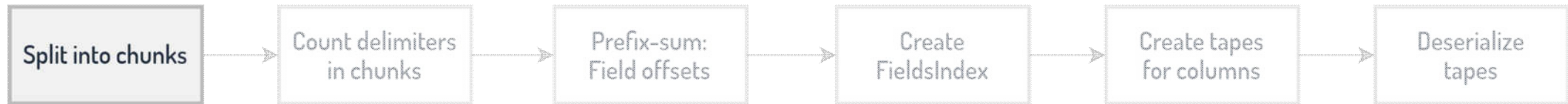


Conceptual Overview: Quoted Mode



Main Challenges

- Partitioning the data into chunks for parallel processing
- Determining each chunk's context
- Deserializing fields in parallel



```
ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 47007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 126150745\n6, China, CN, 1427647786\n7, United States of America, US, 328239523\n8, Canada, CA, 37894799\n...
```

chunk 0
ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 4

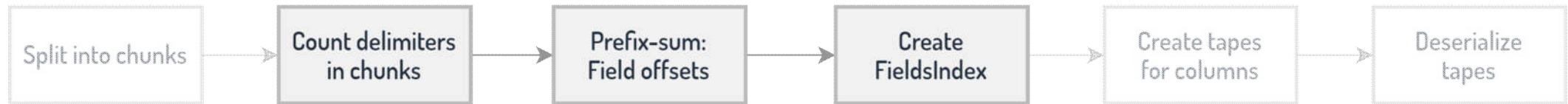
chunk 1
7007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 1261

chunk 2
50745\n6, China, CN, 1427647786\n7, United States of America, US, 328239

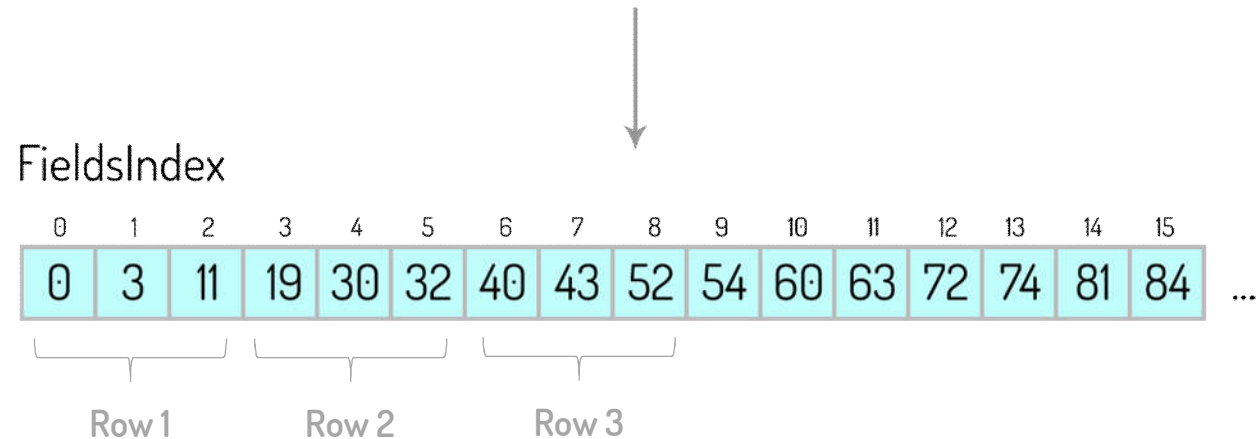
chunk 3
523\n8, Canada, CA, 37894799\n9, Australia, AU, 25721892\n10, Russian F...

...

Load Balancing Warps

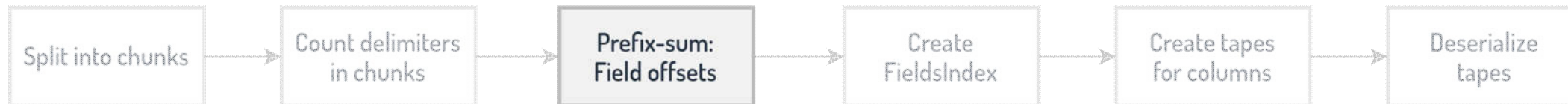


```
ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 47007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 126150745\n6, China, CN, 1427647786\n7, United States of America, US, 328239523\n8, Canada, CA, 37894799\n...
```

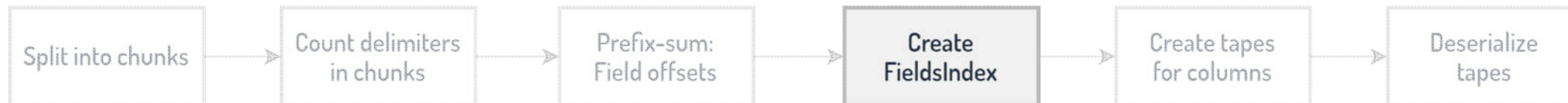




Chunk	Delimiters
chunk 0 ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 4	11
chunk 1 7007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 1261	12
chunk 2 50745\n6, China, CN, 1427647786\n7, United States of America, US, 328239	8
chunk 3 523\n8, Canada, CA, 37894799\n9, Australia, AU, 25721892\n10, Russian F...	10
...	...



Chunk	Delimiters	Prefix sum (offset)
chunk 0 ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 4	11	0
chunk 1 7007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 1261	12	11
chunk 2 50745\n6, China, CN, 1427647786\n7, United States of America, US, 328239	8	23
chunk 3 523\n8, Canada, CA, 37894799\n9, Australia, AU, 25721892\n10, Russian F...	10	31
...	...	...



Prefix sum (offset)		Chunk
0	+	0 1 2 3 4 5 6 7 8 9 10 11 ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 4
11	+	12 13 14 15 16 17 18 19 20 21 22 23 7007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 1261
23	+	24 25 26 27 28 29 30 31 50745\n6, China, CN, 1427647786\n7, United States of America, US, 328239
31	+	32 33 34 35 36 37 38 39 40 41 523\n8, Canada, CA, 37894799\n9, Australia, AU, 25721892\n10, Russian F...
...		...



Prefix sum
(offset)

Chunk

0	+	0 1 2 3 4 5 6 7 8 9 10 11 ID, Country, ISOCode, Population\n1, Germany, DE, 83149319\n2, Spain, ES, 4
11	+	12 13 14 15 16 17 18 19 20 21 22 23 7007367\n3, France, FR, 67076431\n4, Italy, IT, 60317116\n5, Japan, JP, 1261
23	+	24 25 26 27 28 29 30 31 50745\n6, China, CN, 1427647786\n7, United States of America, US, 328239
31	+	32 33 34 35 36 37 38 39 40 41 523\n8, Canada, CA, 37894799\n9, Australia, AU, 25721892\n10, Russian F...
...		...



FieldsIndex

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
0	3	11	19	30	32	40	43	52	54	60	63	72	74	81	84	...

Fast Mode: Quoted Field Delimiters

```
ID,Name,Philosophy\n1,"Aristotle","Quality is not an act, it is a habit."\n2,"Plato","When men speak ill of thee, live so as nobody may believe them."\n3,"Epictetus","It's not what happens to you, but how you react to it that matters."\n...
```

Fast Mode: Quoted Field Delimiters

```
ID,Name,Philosophy\n1,"Aristotle","Quality is not an act, it is a habit."\n2,"Plato","When men speak ill of thee, live so as nobody may believe them."\n3,"Epictetus","It's not what happens to you, but how you react to it that matters."\n...
```

ID	Name	Philosophy
1	"Aristotle"	"Quality is not an act, it is a habit."
2	"Plato"	"When men speak ill of thee, live so as nobody may believe them."
3	"Epictetus"	"It's not what happens to you, but how you react to it that matters."
...	...	...

Fast Mode: Quoted Field Delimiters

```
ID,Name,Philosophy\n1,"Aristotle","Quality is not an act, it is a habit."\n2,"Plato","When men speak ill of thee, live so as nobody may believe them."\n3,"Epictetus","It's not what happens to you, but how you react to it that matters."\n...
```

ID	Name	Philosophy
1	"Aristotle"	"Quality is not an act, it is a habit."
2	"Plato"	"When men speak ill of thee, live so as nobody may believe them."
3	"Epictetus"	"It's not what happens to you, but how you react to it that matters."
...	...	...

ID	Name	Philosophy
1	"Aristotle"	"Quality is not an act
it is a habit."	2	"Plato"
"When men speak ill of thee	live so as nobody may believe them."	3
...	...	...

Quoted Mode

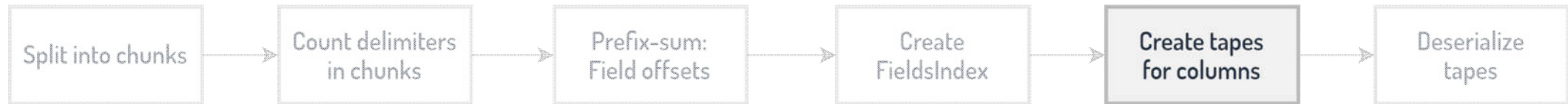
Character is considered *quoted* whenever the number of preceding quotation marks is uneven.

Quoted Mode

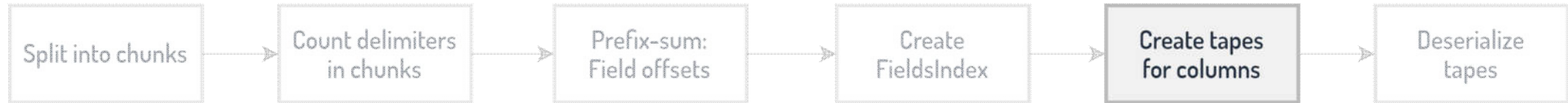
```
ID,Name,Philosophy\n1,"Aristotle","Quality is not an act, it is a habit."\n2,"Plato","When men speak ill of thee, live so as nobody may believe them."\n3,"Epictetus","It's not what happens to you, but how you react to it that matters."\n...
```

FieldsIndex

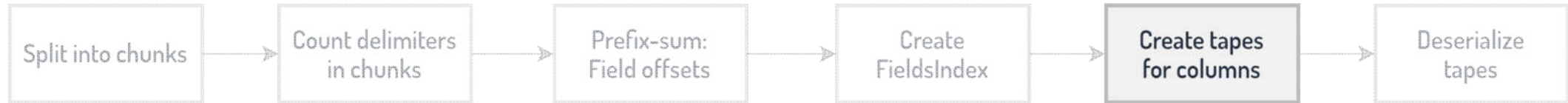




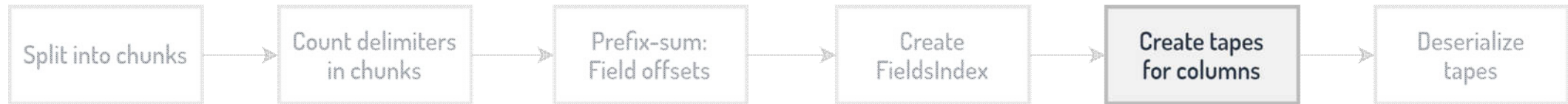
- Deserialization challenges on GPU:
 - Parallelization
 - Warp divergence and occupancy



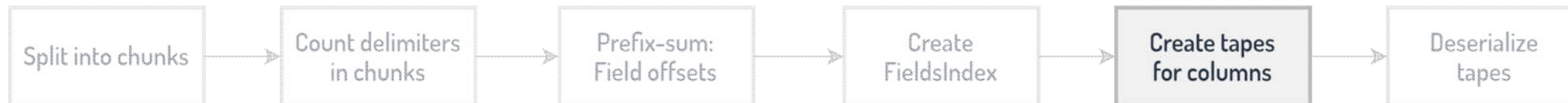
- Deserialization challenges on GPU:
 - Parallelization
 - Warp divergence and occupancy
- Row-based: complex kernels, inefficient memory access



- Deserialization challenges on GPU:
 - Parallelization
 - Warp divergence and occupancy
- Row-based: complex kernels, inefficient memory access
- Column-based: same data type, same kernel logic and control flow



- Like SQL's DDL, users specify a column's max length along its type
`char(15) int(8) float(12) ...`
- Every thread deserializes one field, warp 32 fields in parallel



```

Disease,Year,Country,Cases\n
COVID-19,2020,DE,184492\n
H1N1,2009,ES,1538\n
Spanish Flu,1918,FR,187500\n
...
  
```

Disease : char(12)

C	O	V	I	D	-	1	9	\0	\0	\0	\0
H	1	N	1	\0	\0	\0	\0	\0	\0	\0	\0
S	p	a	n	i	s	h		F	l	u	\0

...

Year : uint(4)

2	0	2	0
2	0	0	9
1	9	1	8

...

Country : char(2)

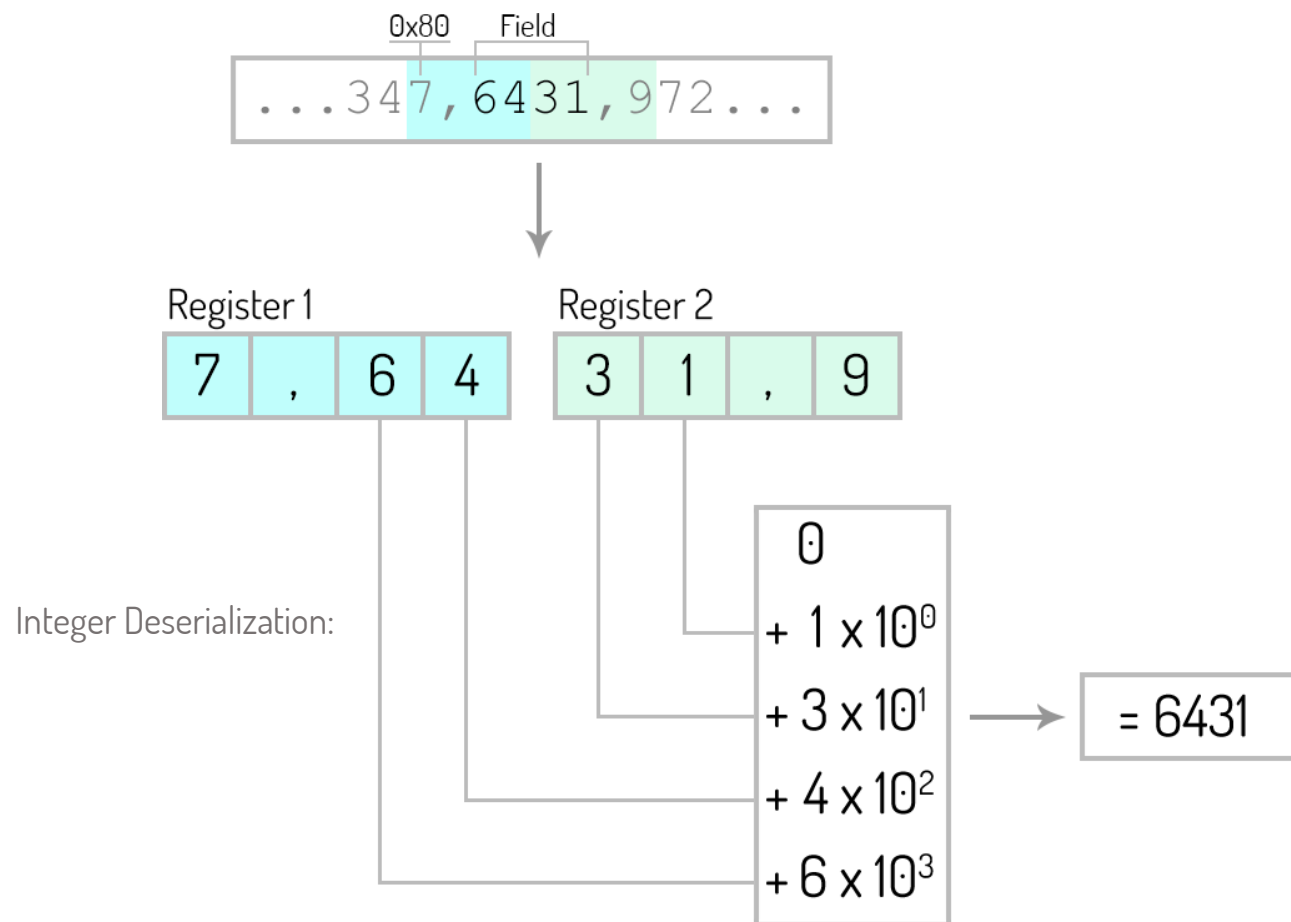
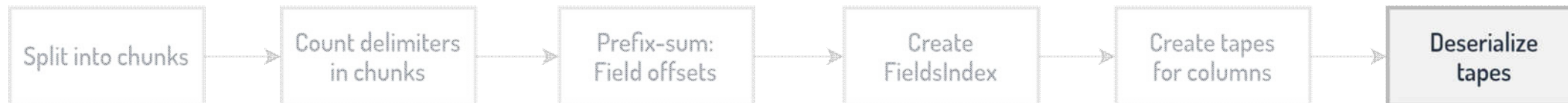
D	E
E	S
F	R

...

Cases : uint(8)

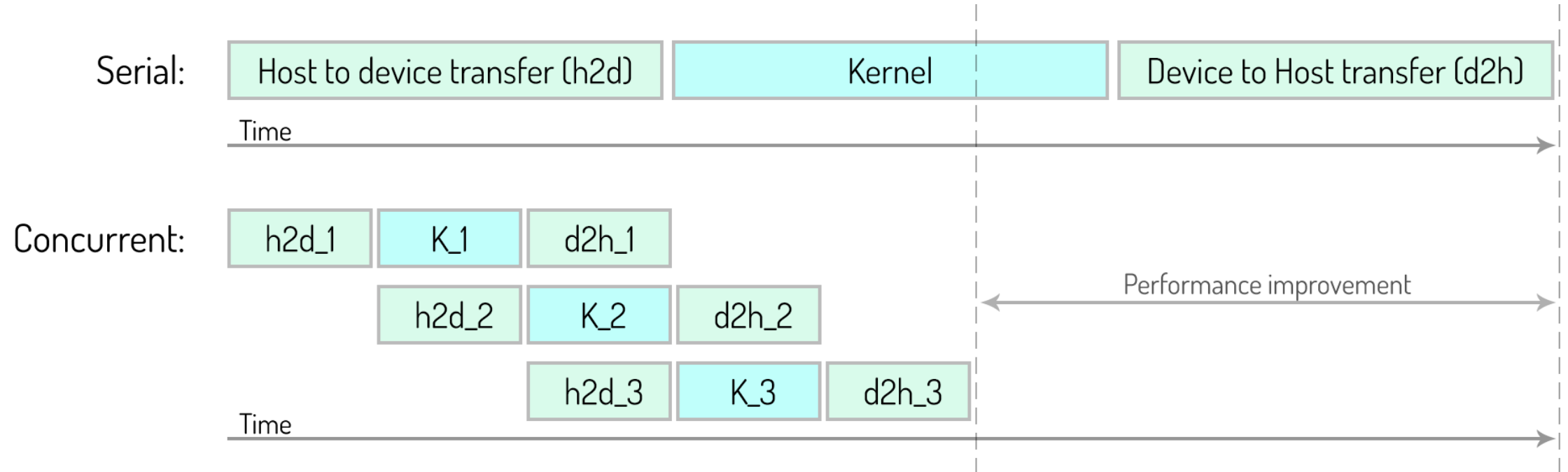
1	8	4	4	9	2	\0	\0
1	5	3	8	\0	\0	\0	\0
1	8	7	5	0	0	\0	\0

...

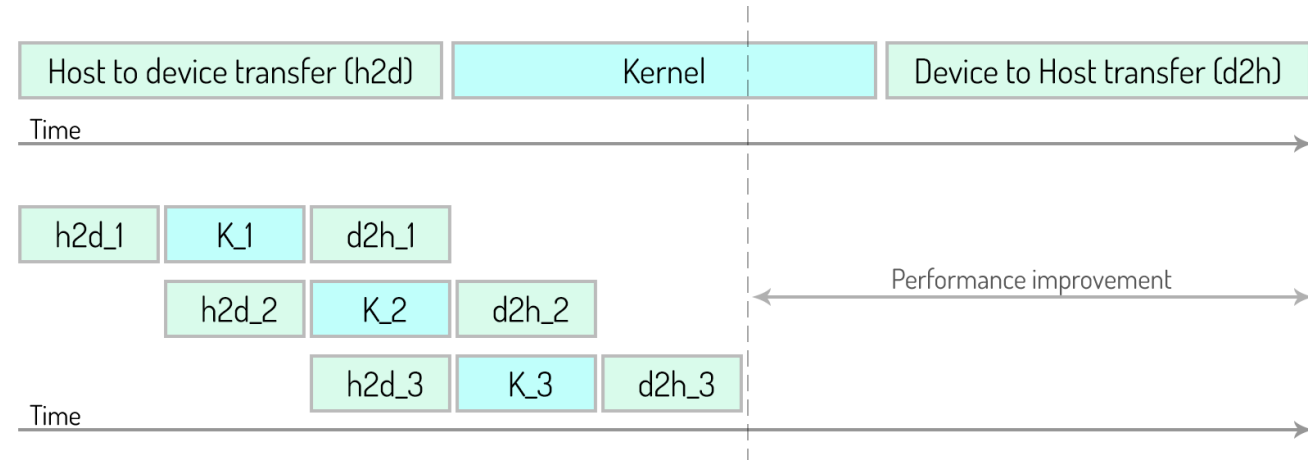
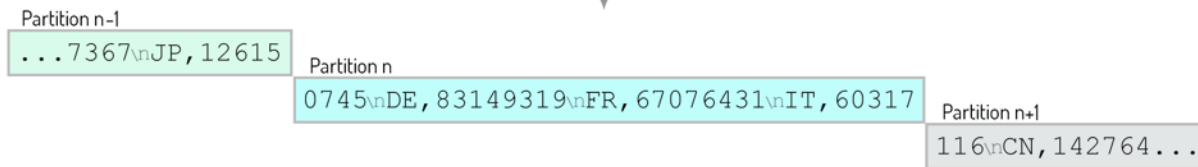
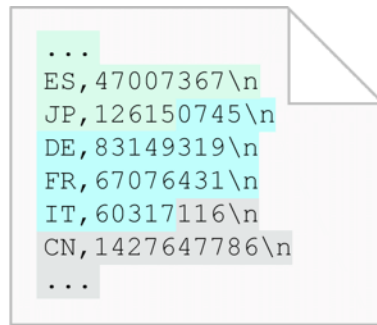


Streaming

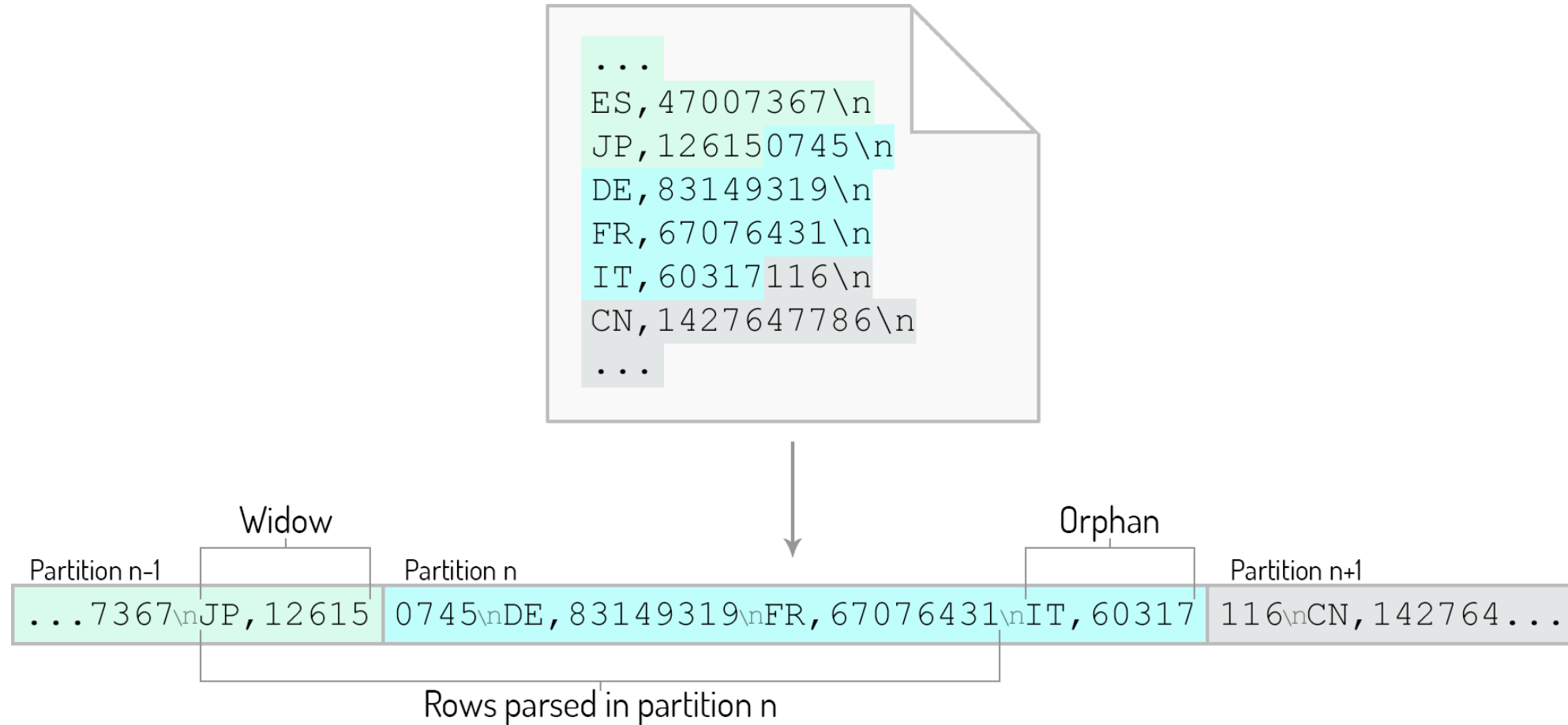
Streaming



Partitions

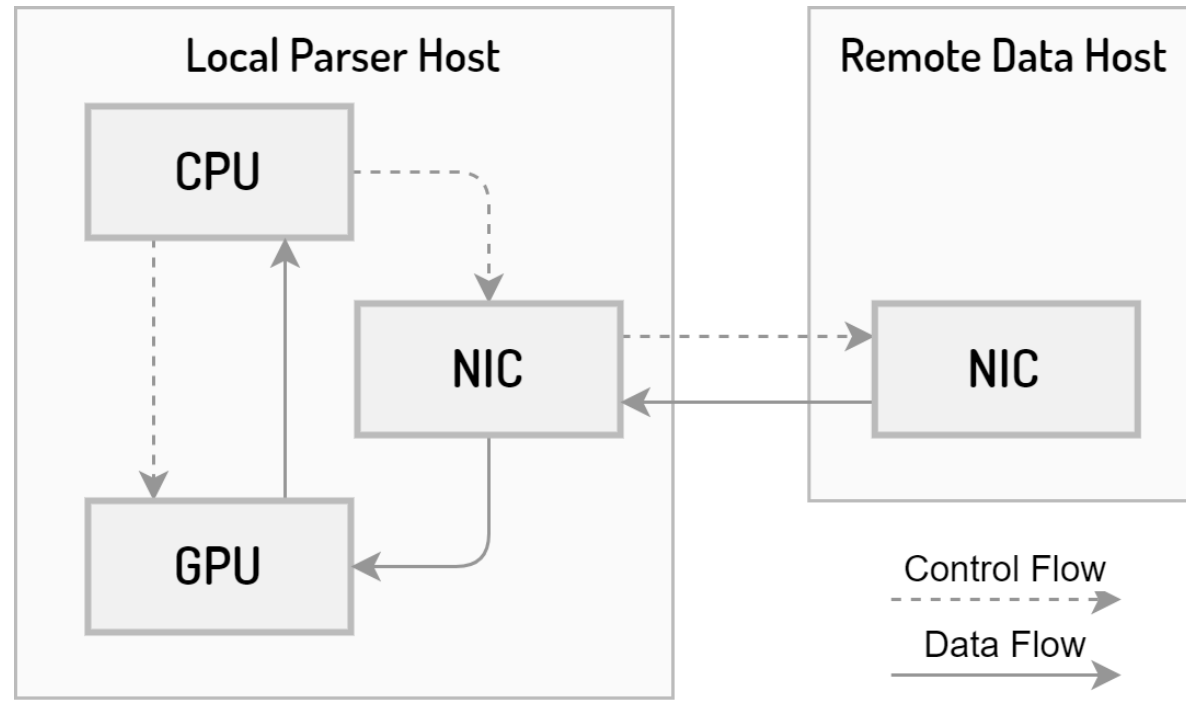


Context Handover



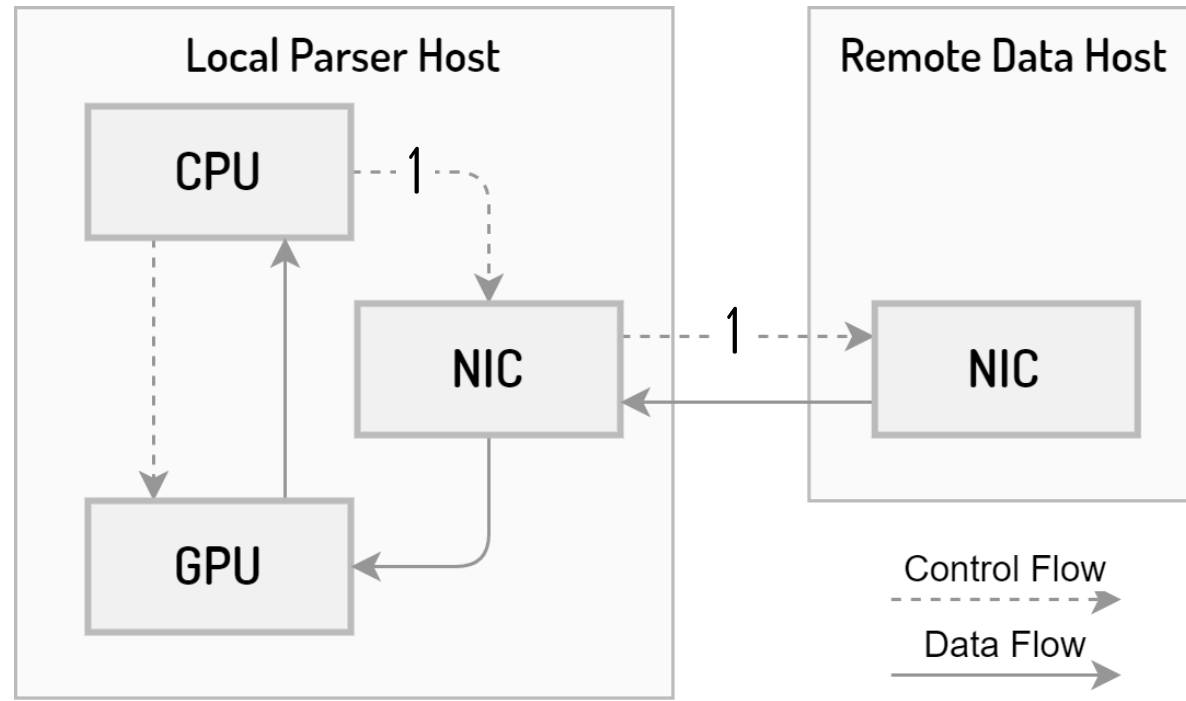
End-to-End Loading with RDMA

Partition assigned to free worker:



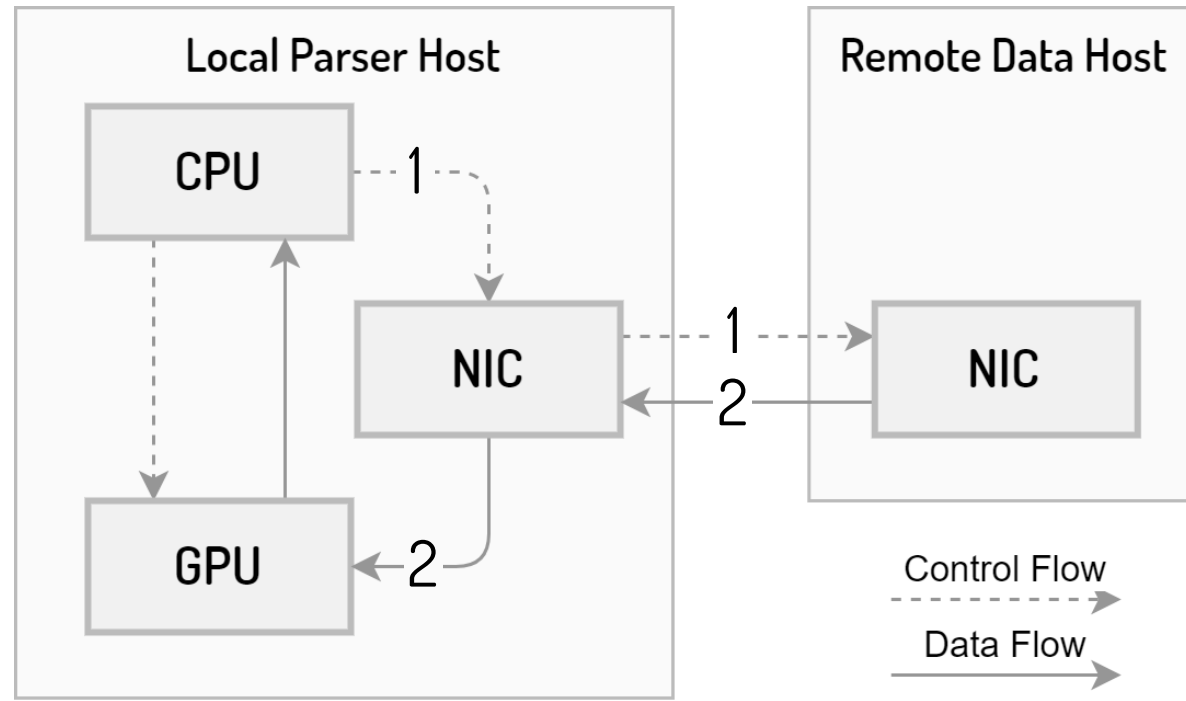
End-to-End Loading with RDMA

Partition assigned to free worker:



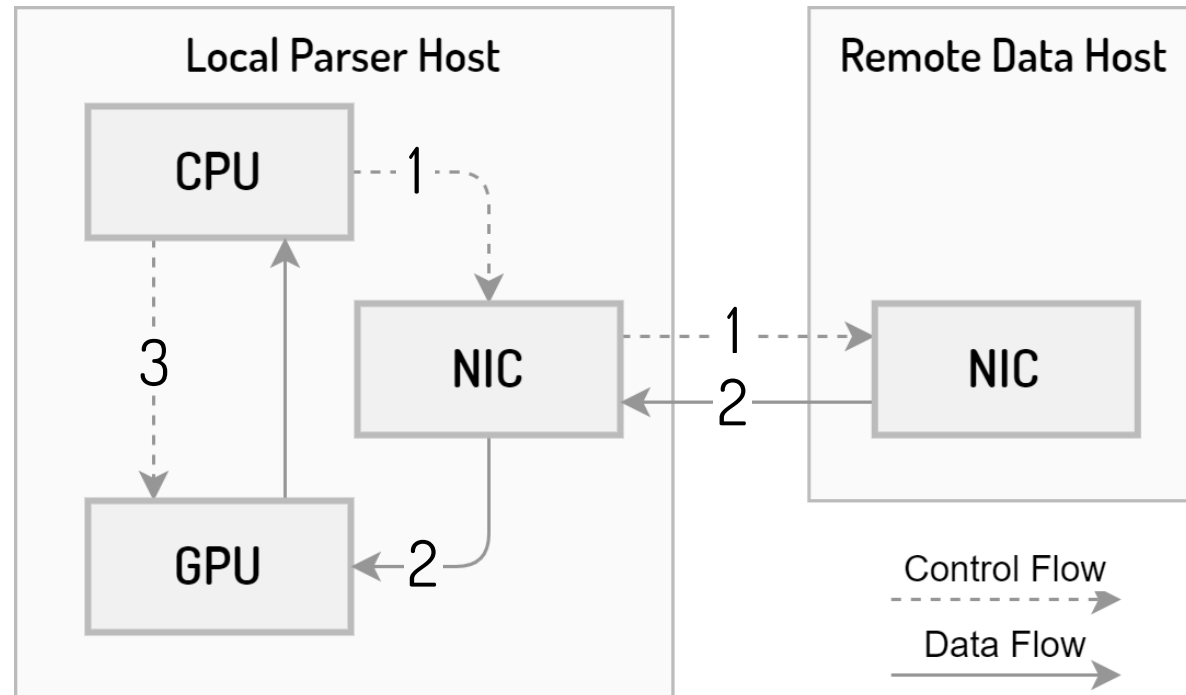
End-to-End Loading with RDMA

Partition assigned to free worker:



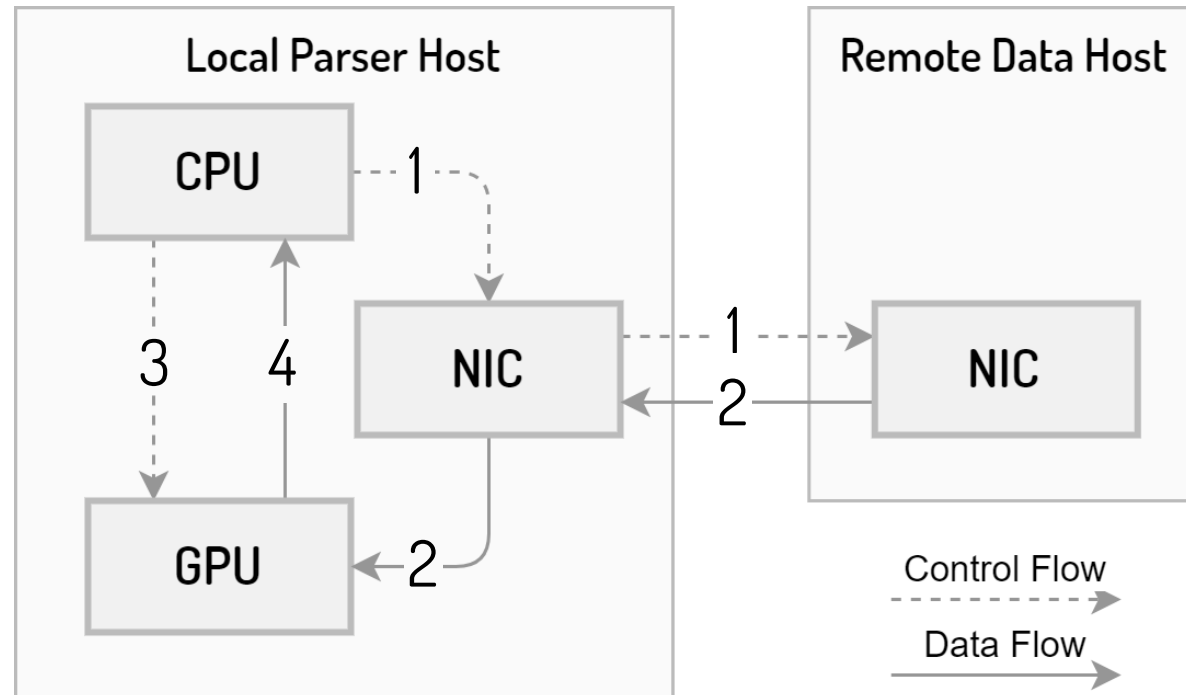
End-to-End Loading with RDMA

Partition assigned to free worker:



End-to-End Loading with RDMA

Partition assigned to free worker:



Evaluation

Experiment Setup

- Intel Xeon Gold 5115 (10c/20t)
- Nvidia Tesla V100 (PCIe 3.0)
- Nvidia Tesla V100 (NVLink 2.0)
- InfiniBand 100 Gb/s
- NYC Yellow Taxi Trips
(1.9 GB, 22.5M rows, 18 cols)
- TPC-H Lineitem
(719 MB, 6M rows, 16 cols)
- int_444
(1 GB, 70M rows, 3 columns)

CPU vs. GPU

NYC Yellow Taxi

THROUGHPUT

30 GB/s

25 GB/s

20 GB/s

15 GB/s

10 GB/s

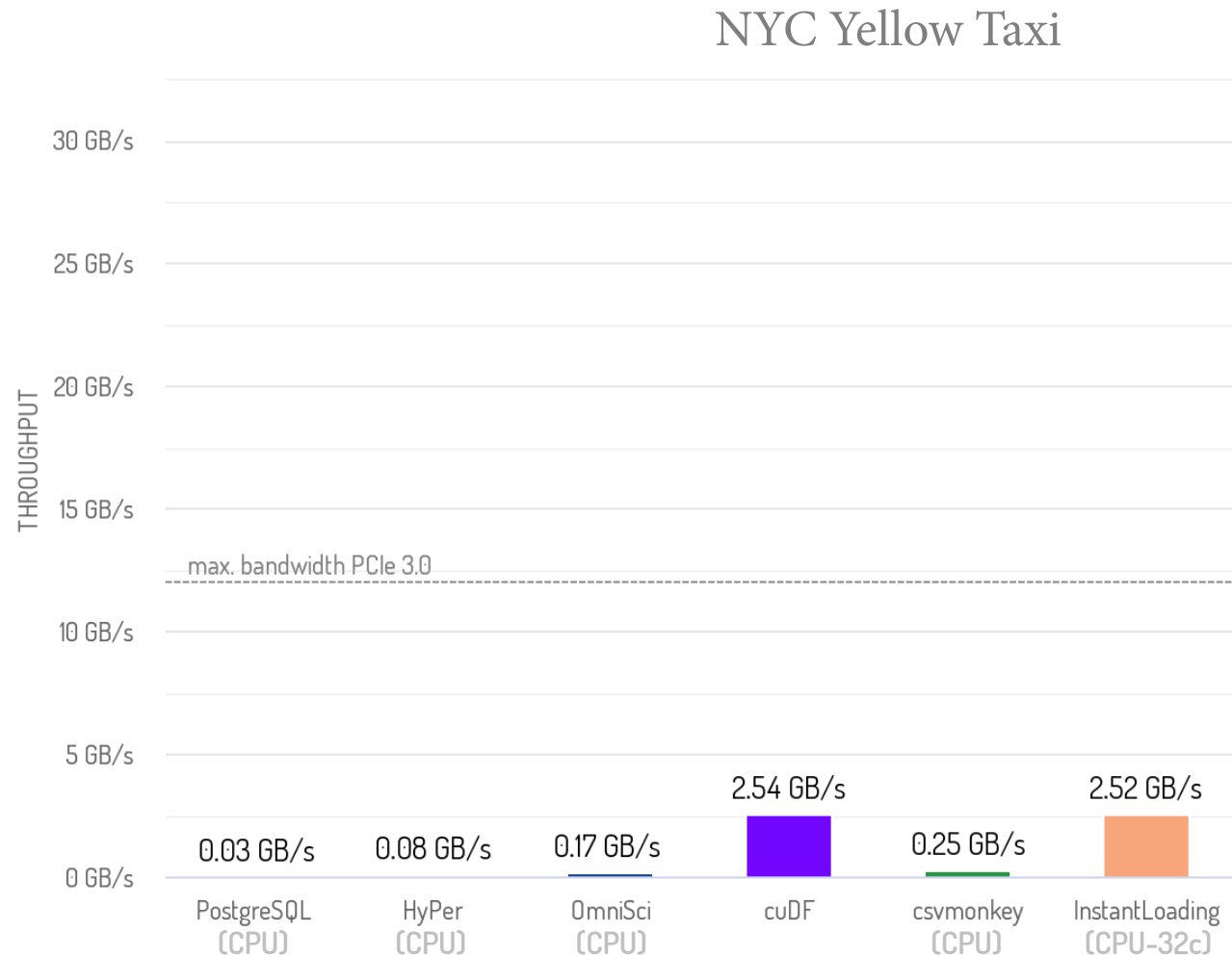
5 GB/s

0 GB/s

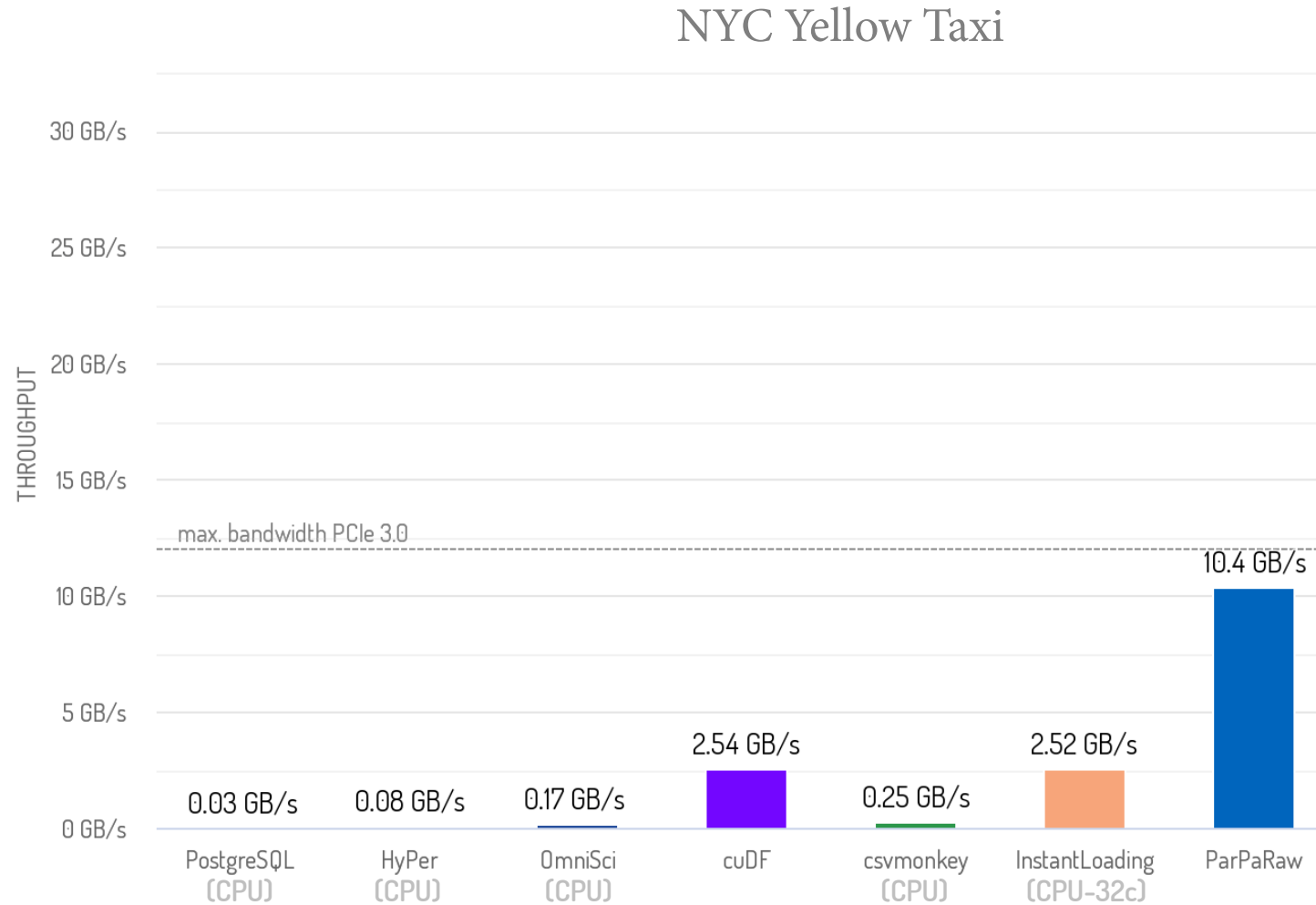
CPU vs. GPU



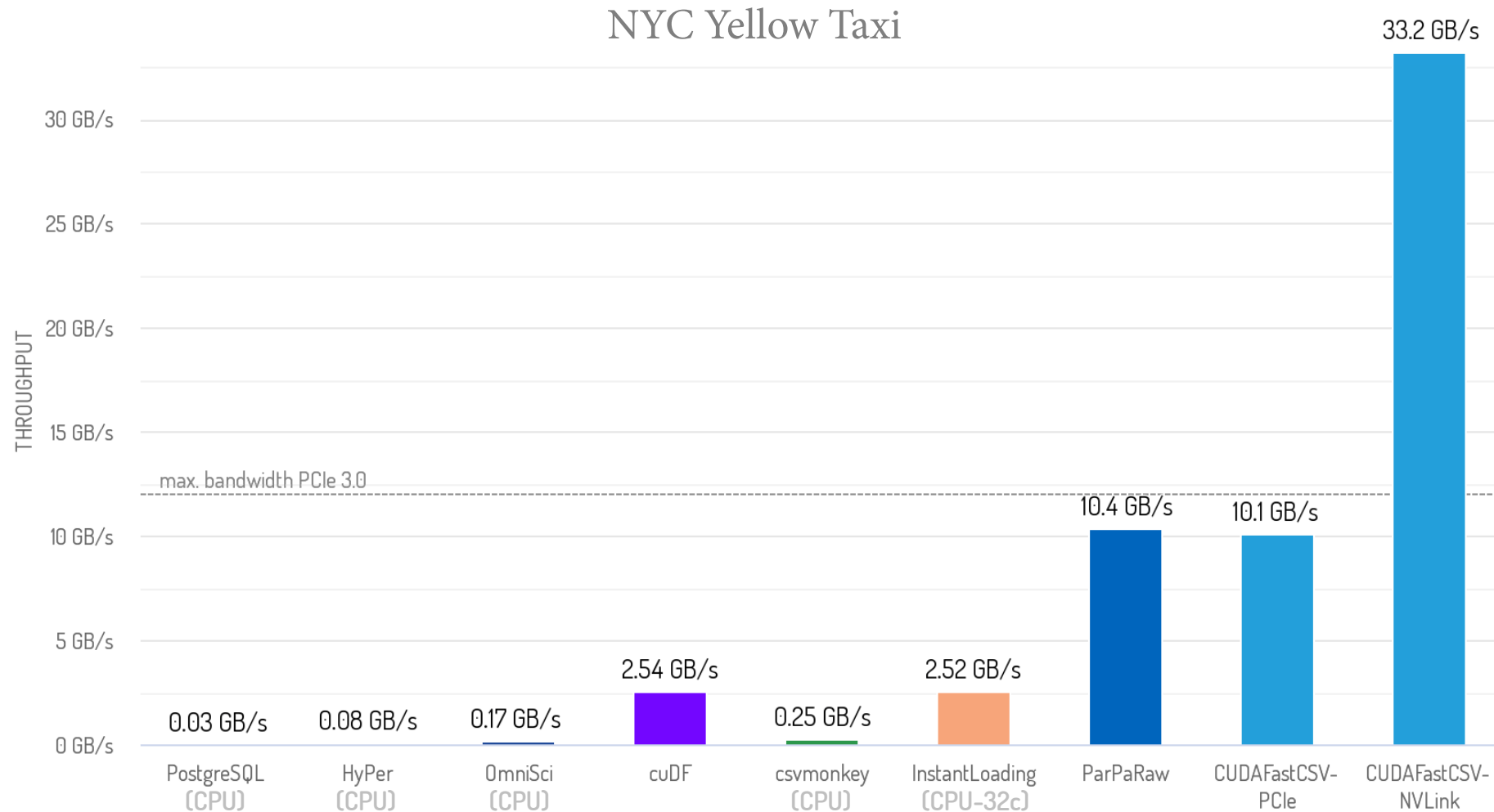
CPU vs. GPU



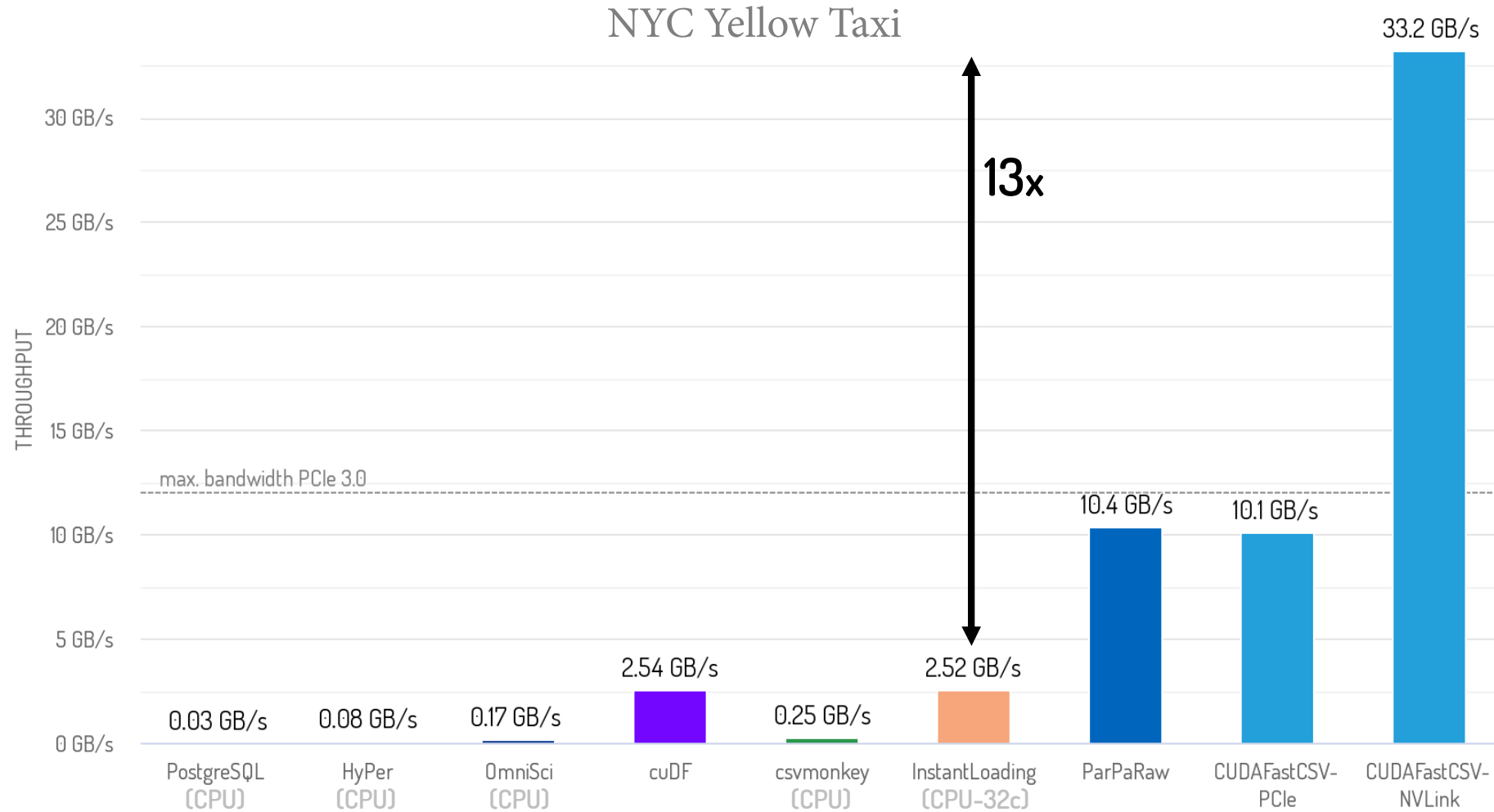
CPU vs. GPU



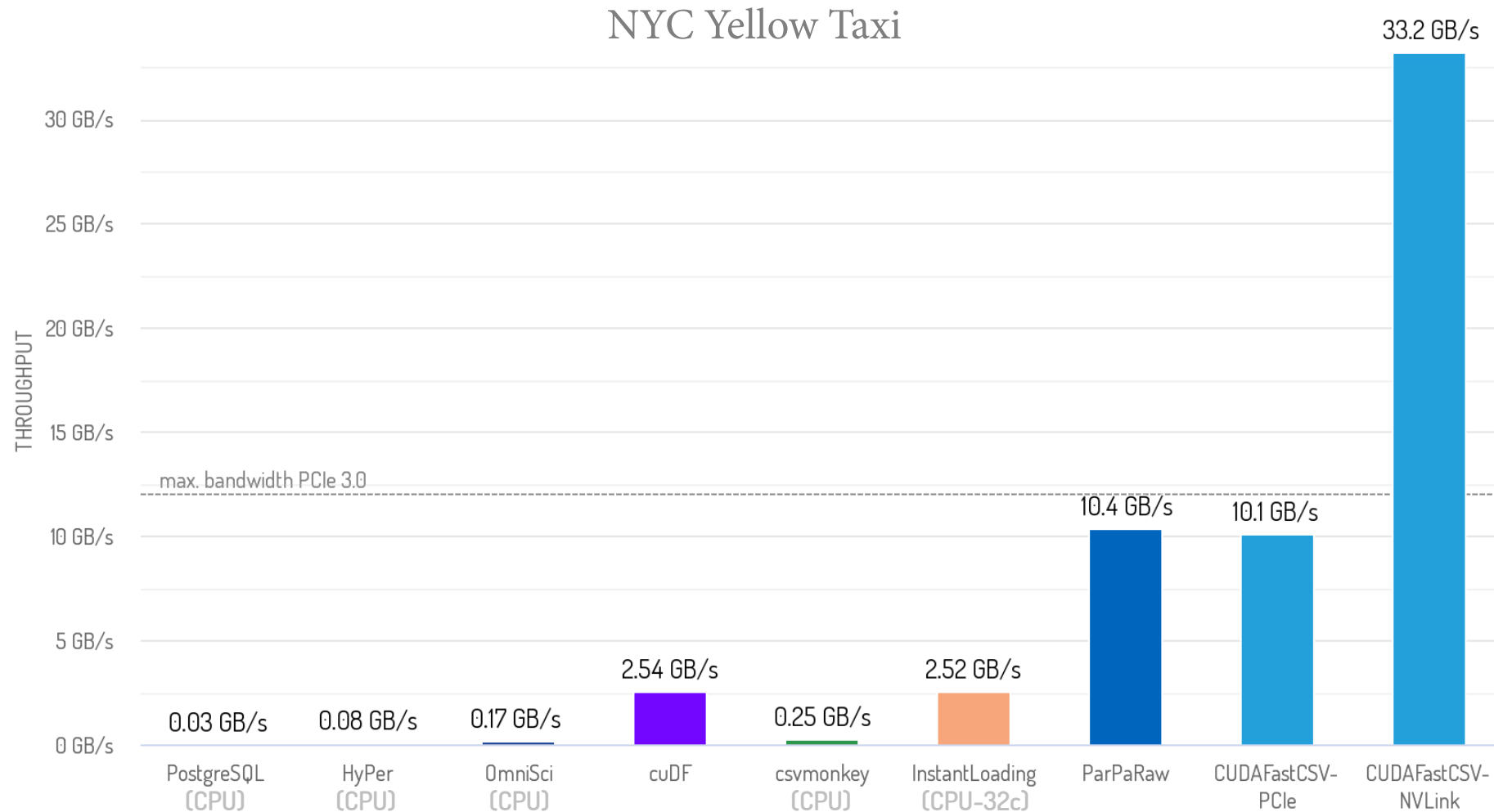
CPU vs. GPU



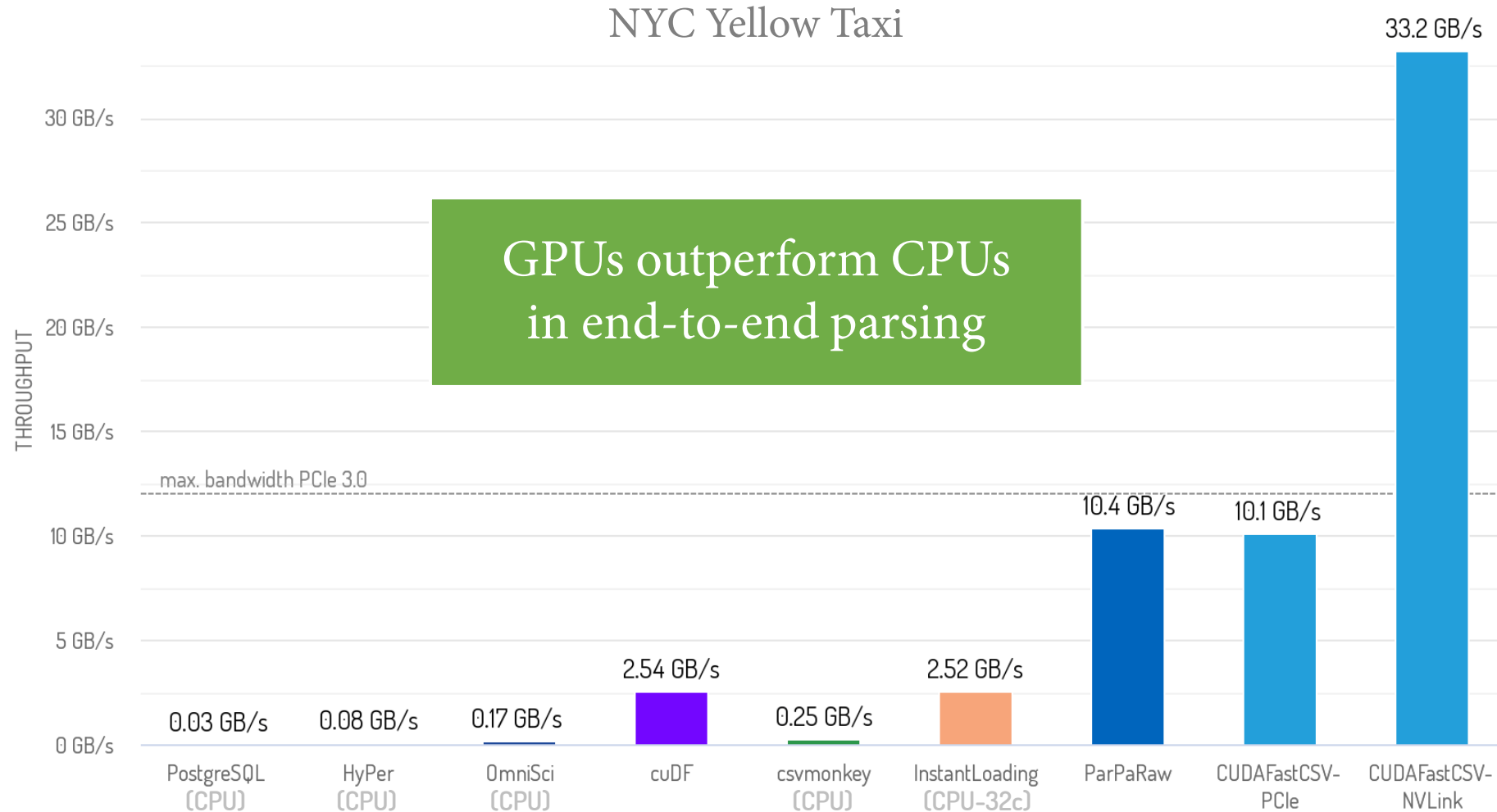
CPU vs. GPU



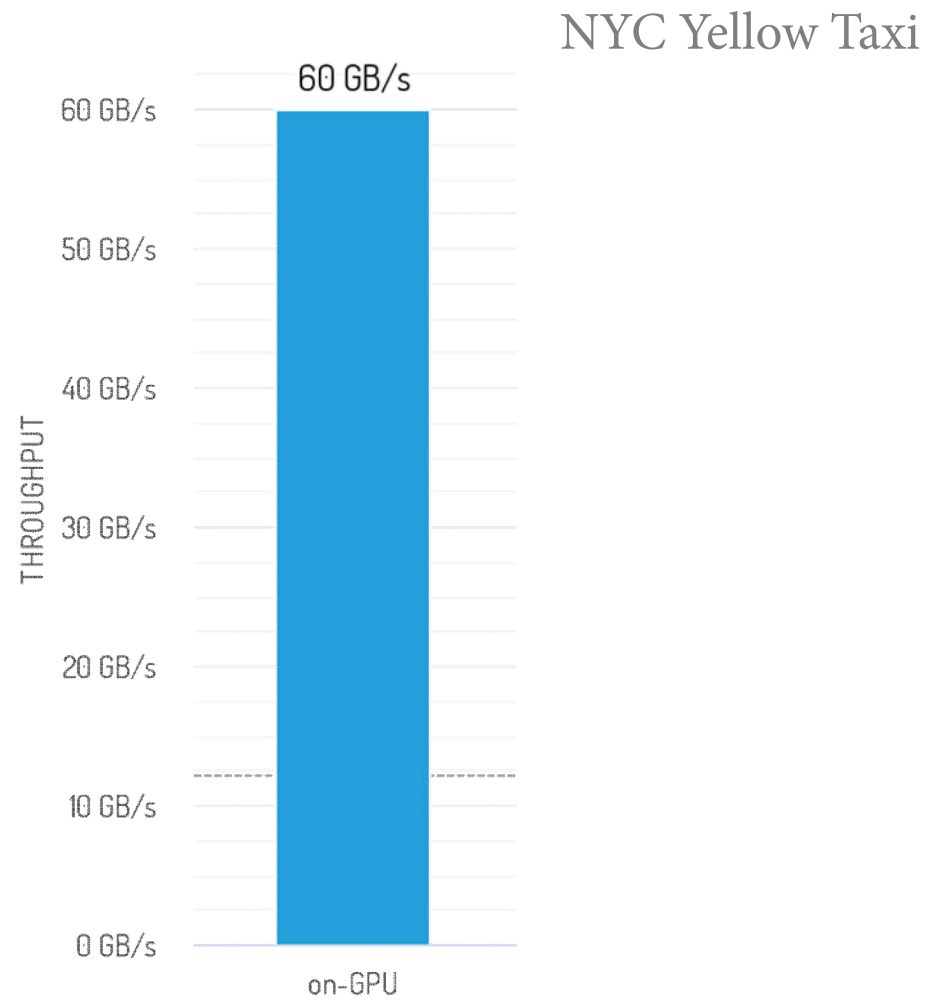
CPU vs. GPU



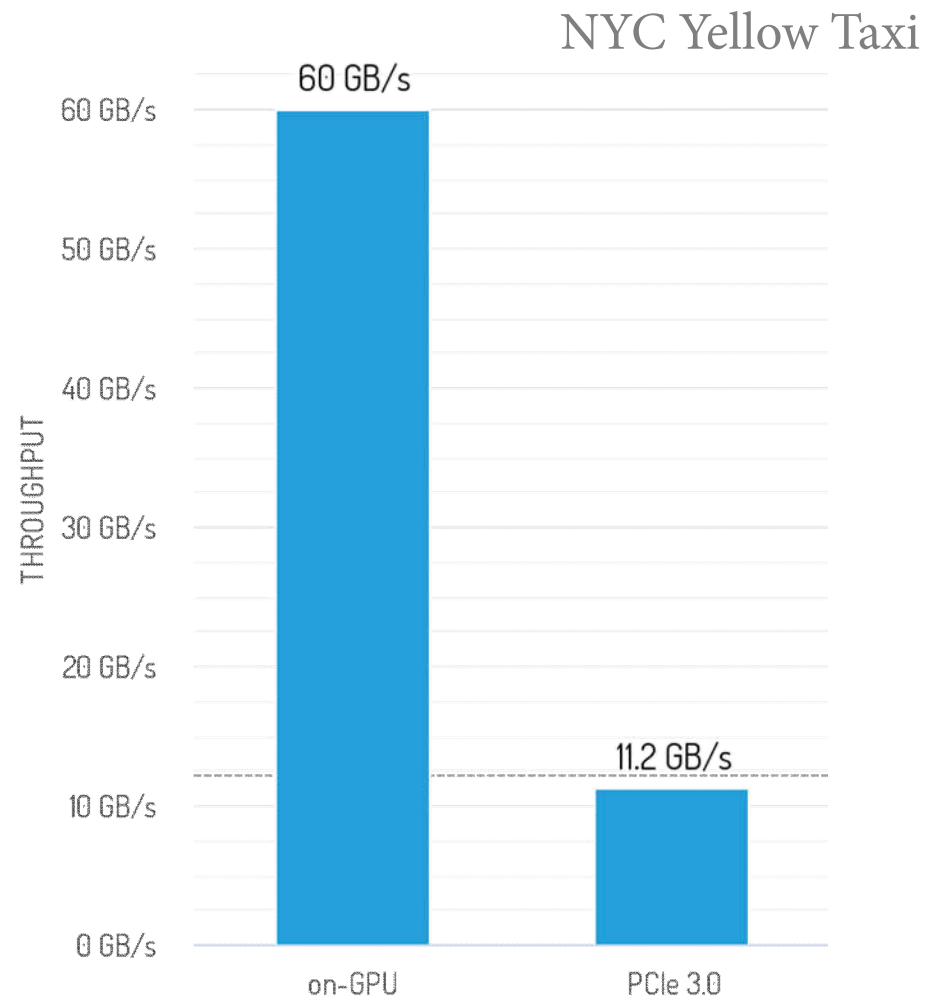
CPU vs. GPU



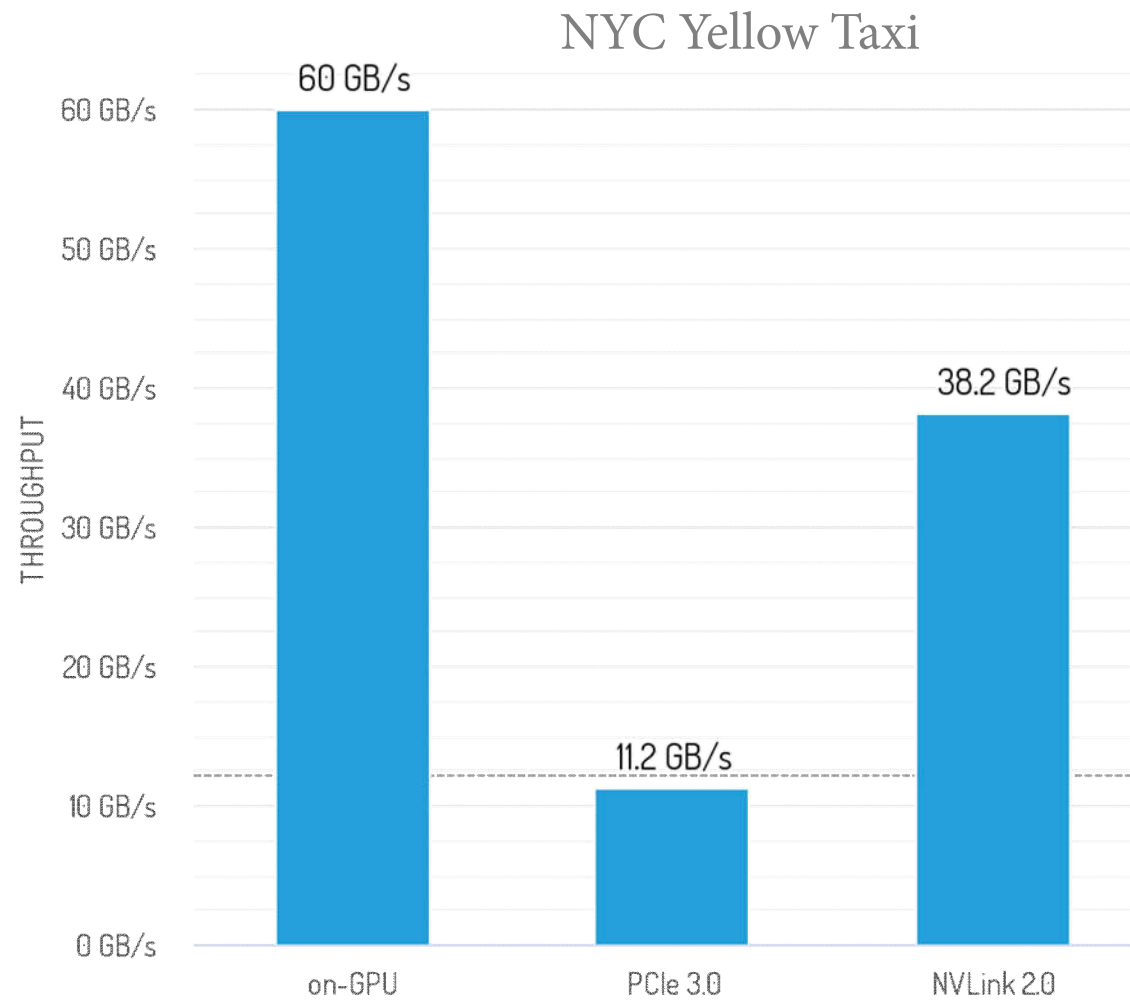
Streaming I/O



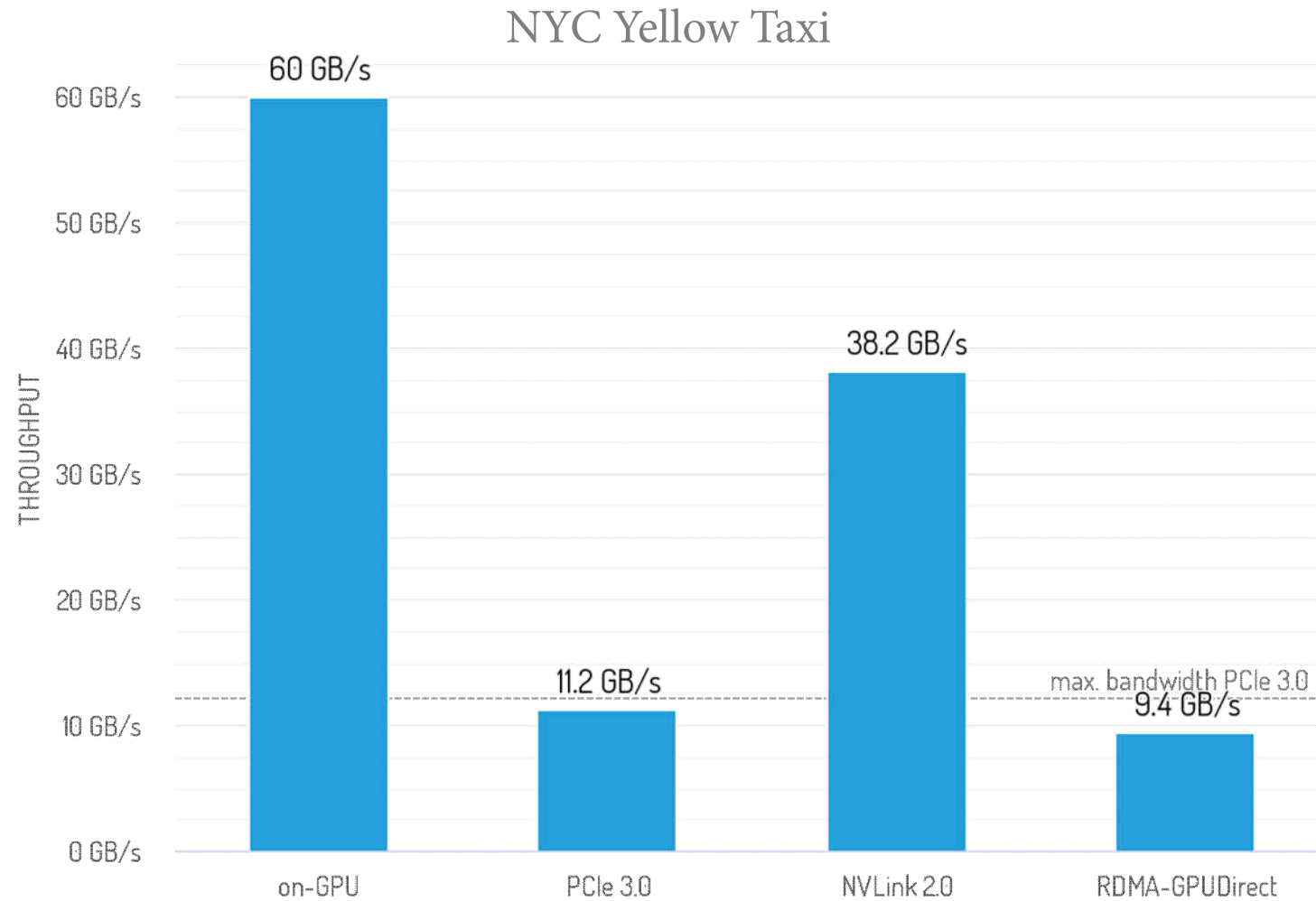
Streaming I/O



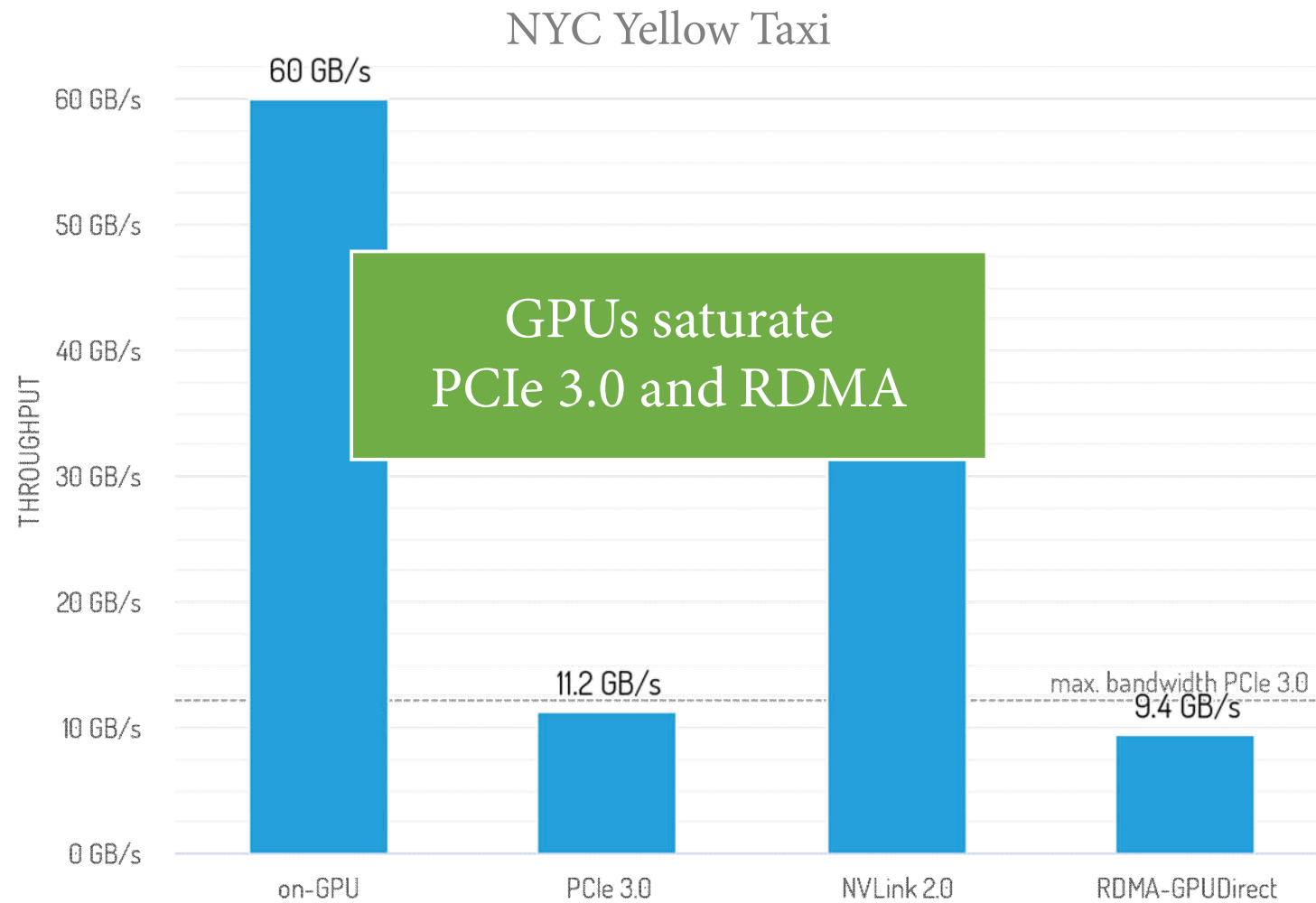
Streaming I/O



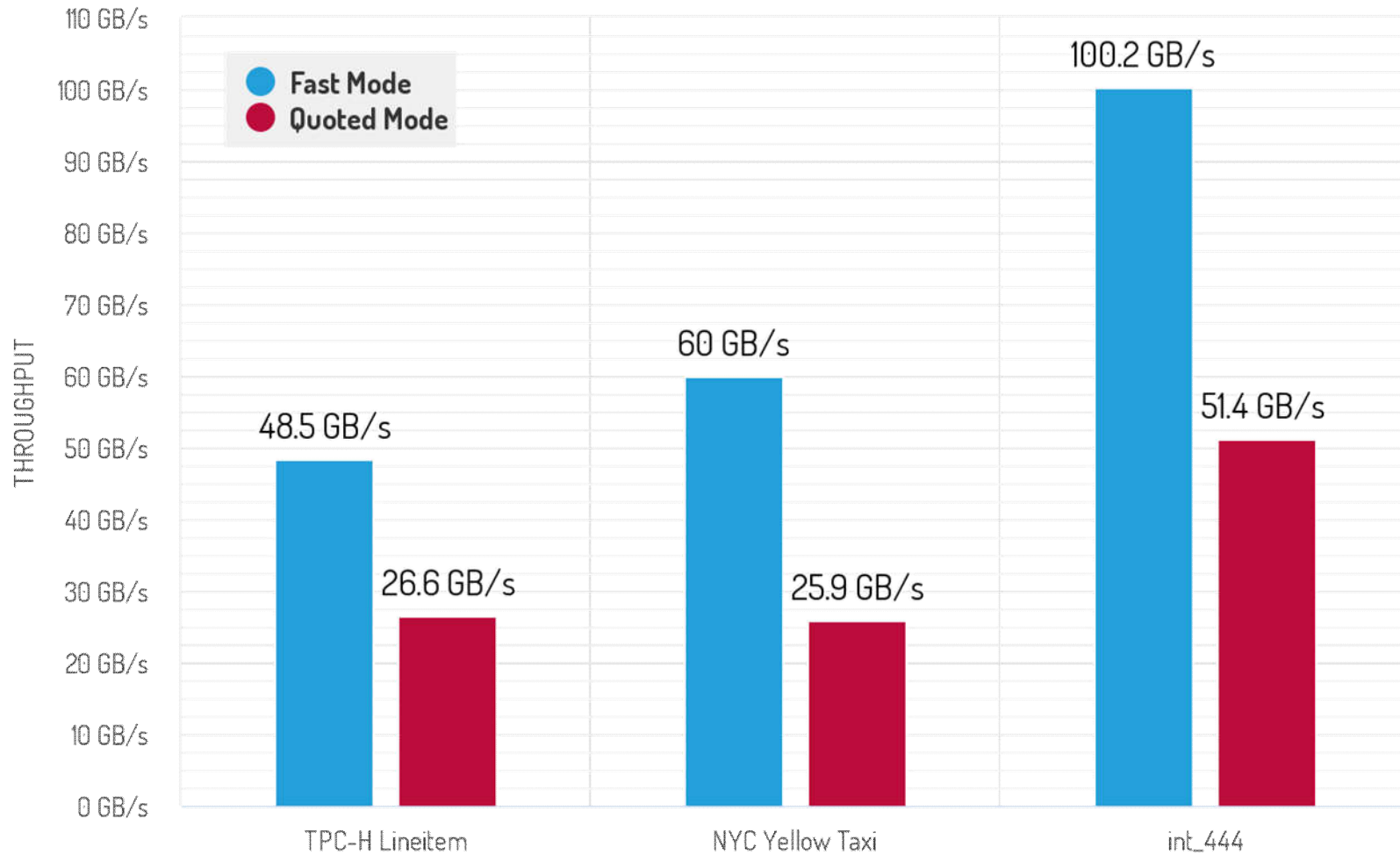
Streaming I/O



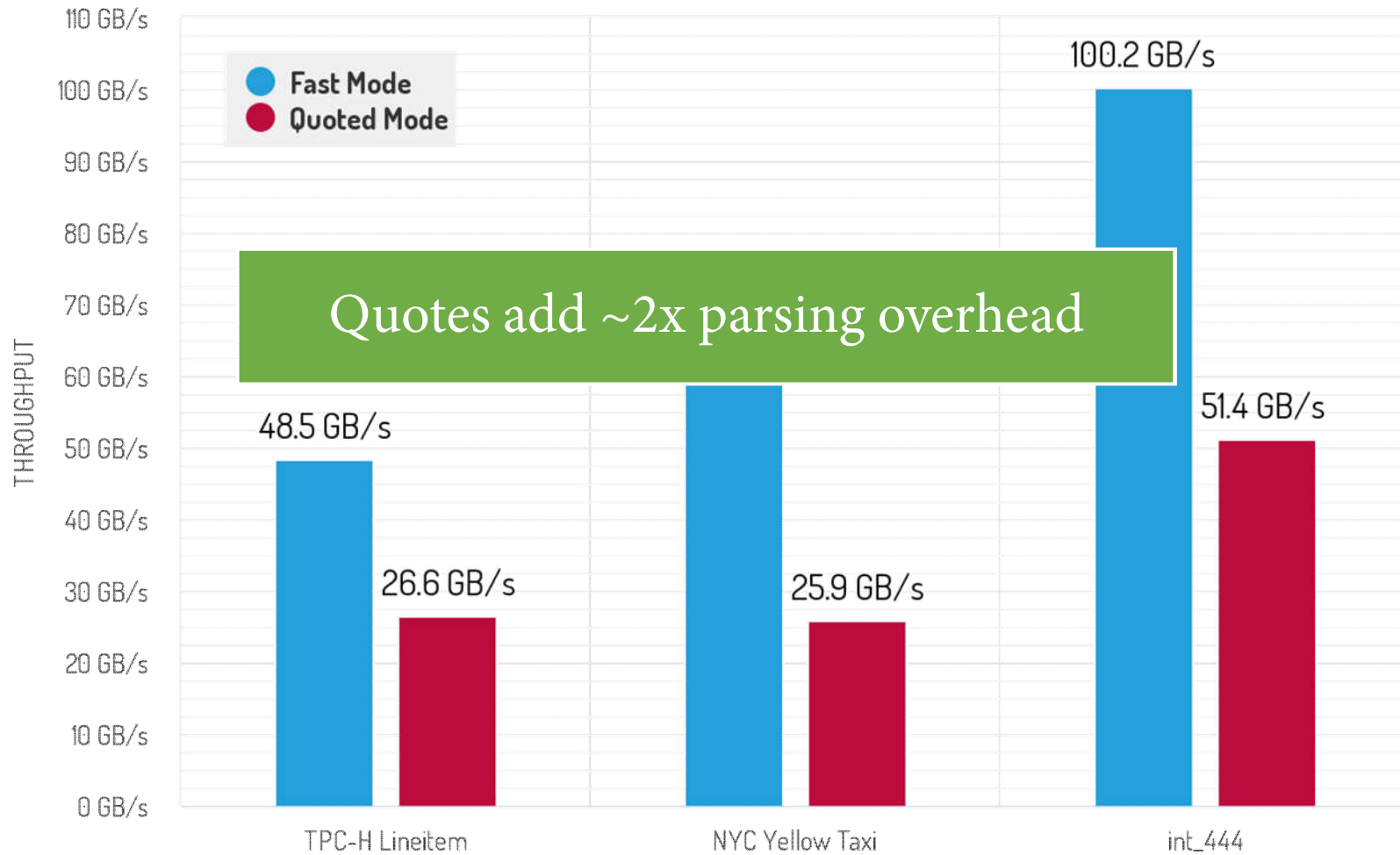
Streaming I/O



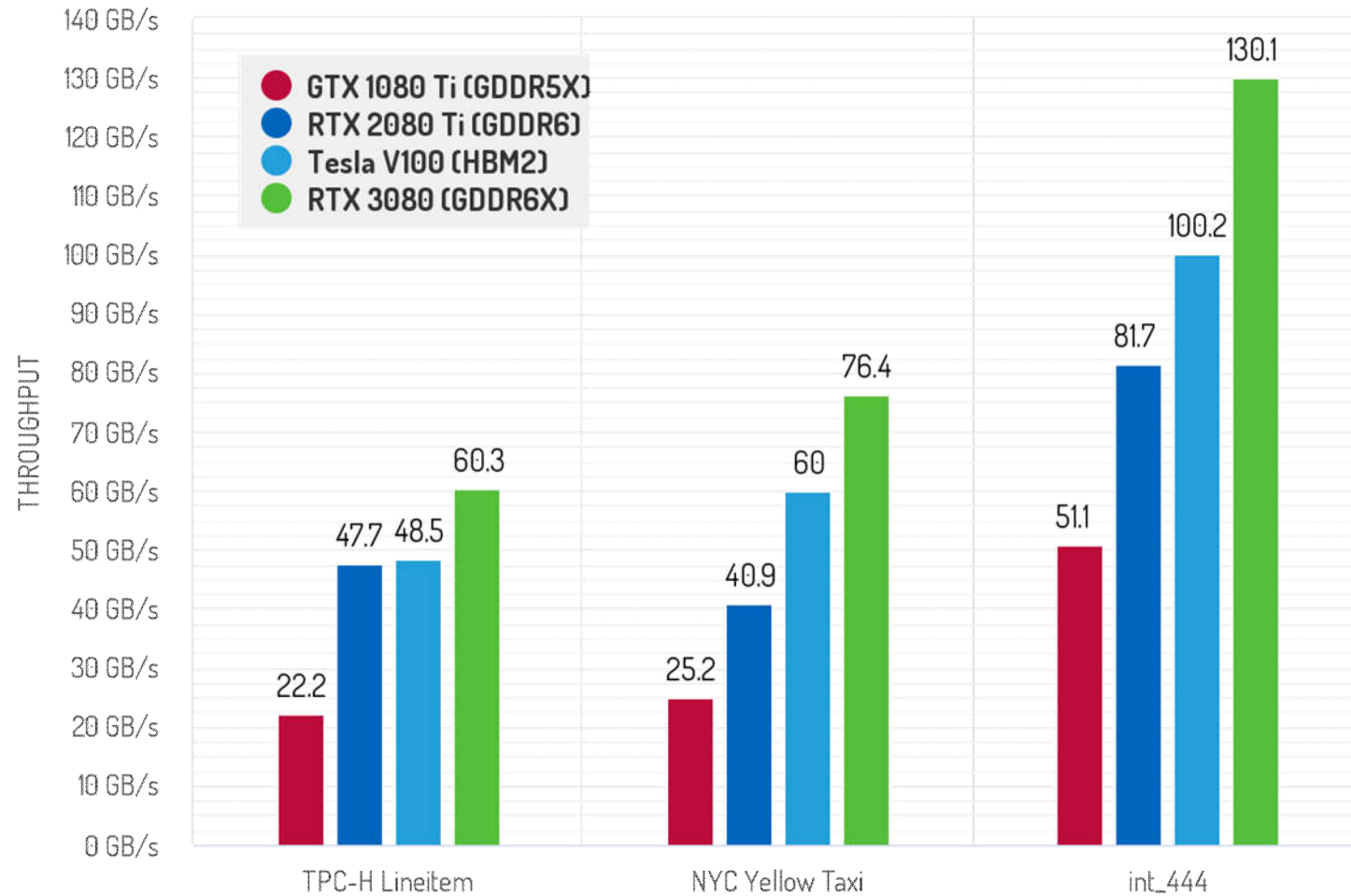
Fast Mode vs. Quoted Mode



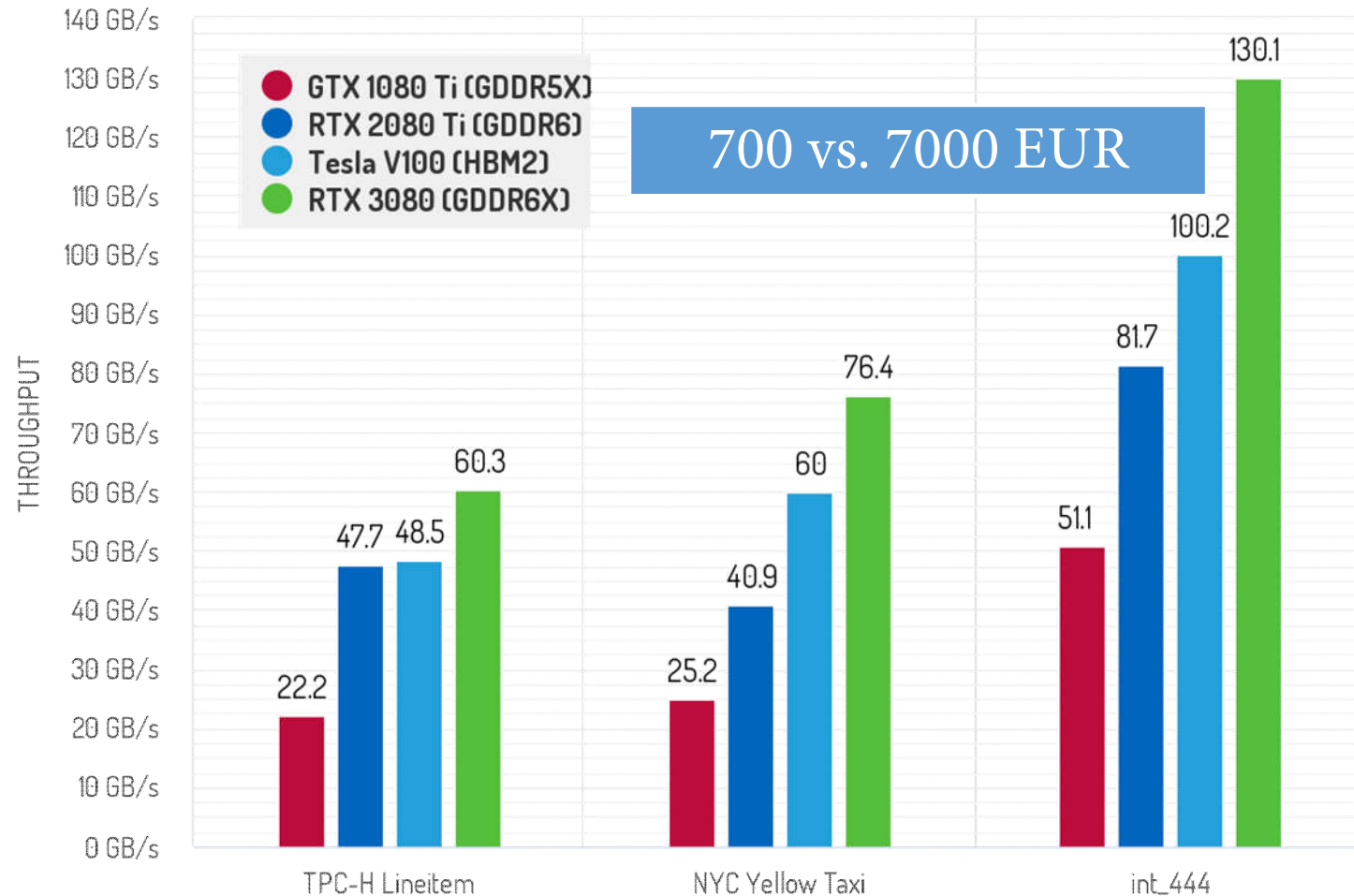
Fast Mode vs. Quoted Mode



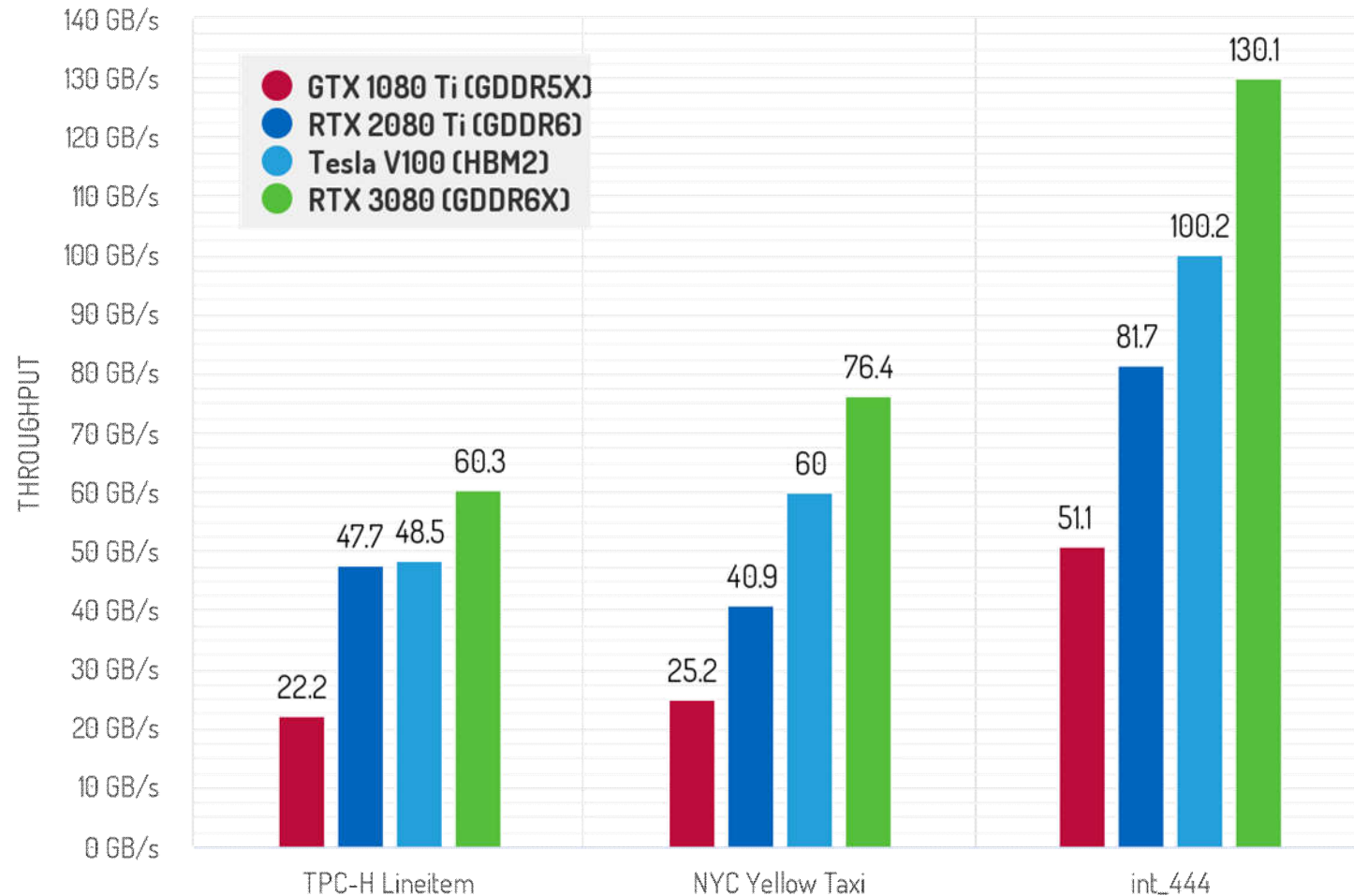
Hardware Scalability



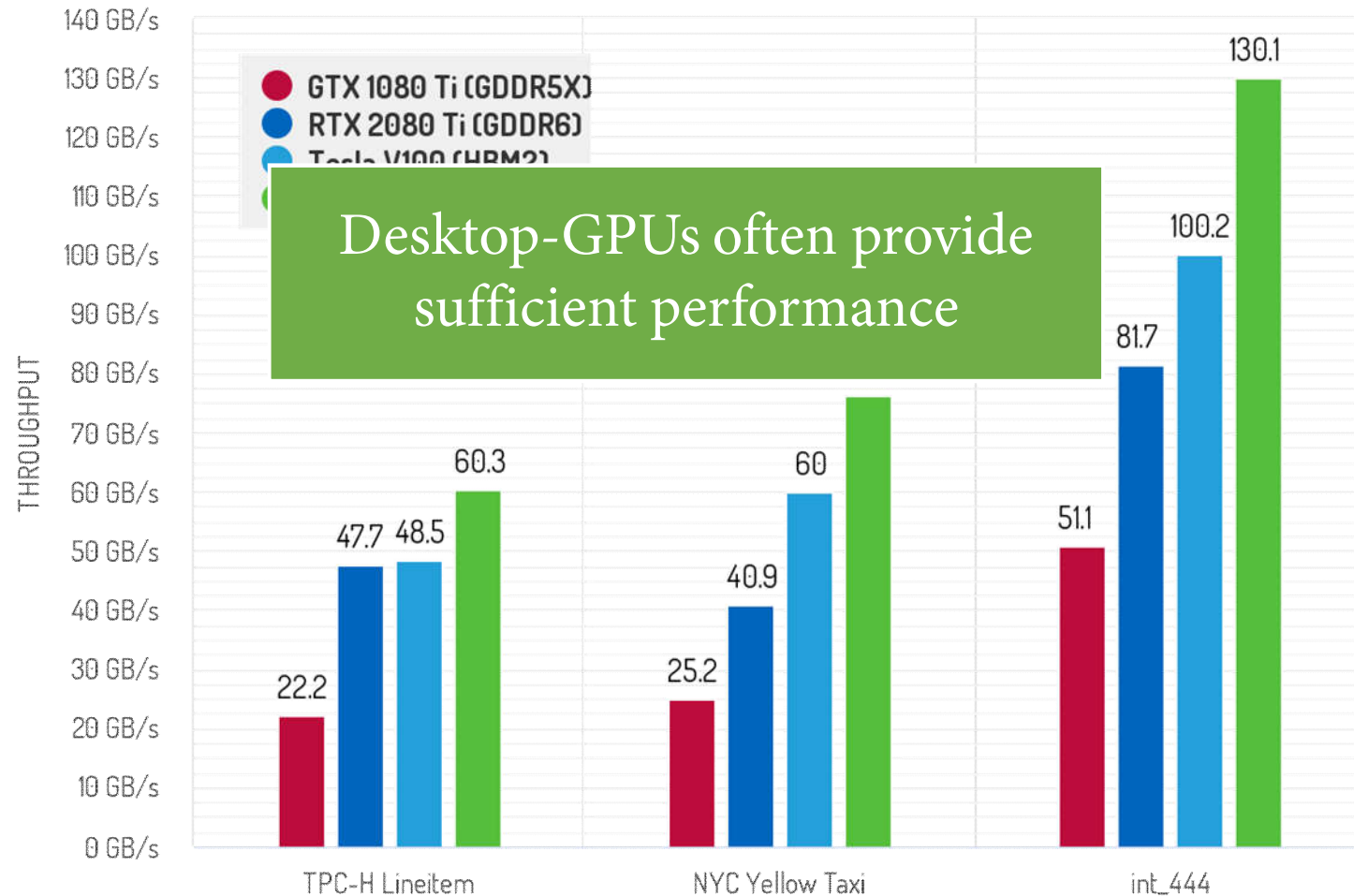
Hardware Scalability



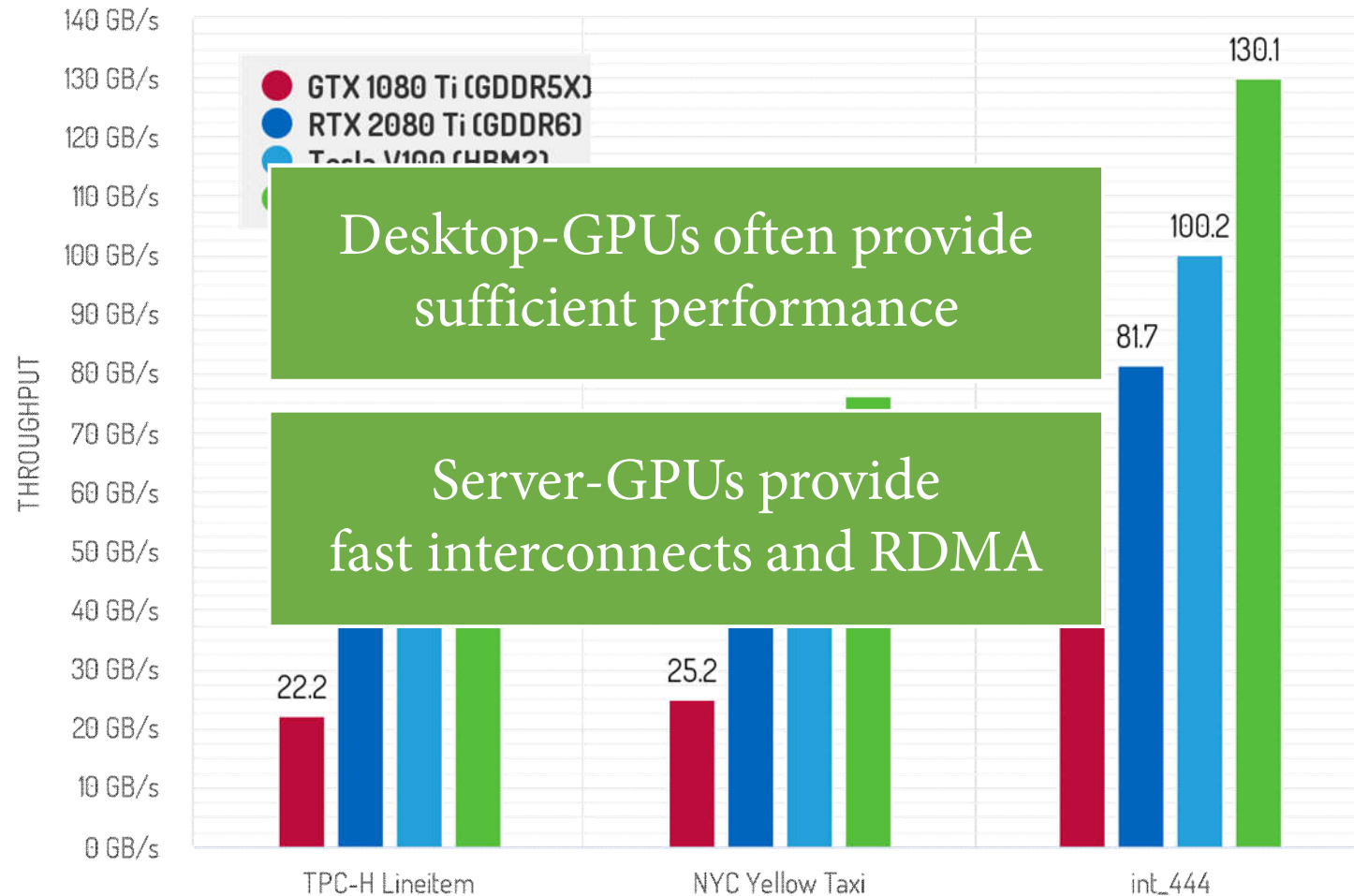
Hardware Scalability



Hardware Scalability



Hardware Scalability



Conclusion

- GPUs can solve data loading bottleneck
- New way to integrate GPUs into DBs