Liberal throughput estimates: to prove GPU isn't worthwhile for FP64 linalg					
	FADD	FMUL	FFMA	FRECIP	FRSQRT
Single Precision (e8m24) GIPS					
M1 1-CPU	51.65	51.65	51.65	12.91	12.91
M1 8-CPU	388.61	388.61	388.61	97.15	97.15
M1 4-AMX*	0.00	0.00	826.37	0.00	0.00
M1 8-AMX*	0.00	0.00	1652.74	0.00	0.00
M1 32-GPU	5308.42	5308.42	5308.42	884.74	663.55
i9-13900K 24-CPU	1792.00	1792.00	1792.00	179.20	179.20
RTX 4090 128-GPU	41287.68	41287.68	41287.68	5160.96	5160.96
CPU/core/GHz	16.00	16.00	16.00	4.00	4.00
AMX/core/GHz	0.00	0.00	64.00	0.00	0.00
M1/core/GHz	128.00	128.00	128.00	21.33	16.00
RTX/core/GHz	128.00	128.00	128.00	16.00	16.00
*1-CPU, 4-AMX, 8-AMX assumed 3.228 GHz. 8-CPU assumed 3.036 GHz					
**Assuming all P/E-cores reach turbo boost, 2x AVX-256 per cycle					
Single-Double (e8m48) GIPS					
M1 32-GPU	482.73	663.55	279.52	94.97	106.17
RTX 4090 128-GPU	3548.16	4193.28	1922.46	-	-
M1/core/GHz	11.64	16.00	6.74	2.29	2.56
RTX/core/GHz	11.00	13.00	5.96	-	-
Double Precision (e11m48) GIPS					
M1 1-CPU	25.82	25.82	25.82	6.46	6.46
M1 8-CPU	194.30	194.30	194.30	48.58	48.58
M1 4-AMX	0.00	0.00	206.59	0.00	0.00
M1 8-AMX	0.00	0.00	413.18	0.00	0.00
M1 32-GPU	147.64	147.64	100.36	77.97	85.43
i9-13900K 24-CPU	896.00	896.00	896.00	56.00	56.00
RTX 4090 128-GPU	645.12	645.12	645.12	161.28	161.28
CPU/core/GHz	8.00	8.00	8.00	2.00	2.00
AMX/core/GHz	0.00	0.00	16.00	0.00	0.00
M1/core/GHz	3.56	3.56	2.42	1.88	2.06
RTX/core/GHz	2.00	2.00	2.00	0.50	0.50
Operation	FADD	FMUL	FFMA	FRECIP	FRSQRT
CADD	2	0	0	0	0
CMUL	0	2	2	0	0
CFMA	0	0	4	0	0
CRECIP	0	3	1	1	0
CDIV	0	5	3	1	0
Linalg operation, num ops					
DGEMV	0	0	n^2	0	0
DGEMM	0	0	n^3	0	0
DPOTRF	0	0	(1/6)n^3	(1/2)n^2	n

Lower bound, e-FP64 cycles	Complex cycles	Scheduler* cycle	ILP=4 Cycles	ILP=1 Cycles
e11m48 + e11m48	24 + 8+4	16+2	36	36
e11m48 x e11m48	40 + 8+4	16+2	52	52
e11m48 x e11m48 + e11m48	52 + 12+4	24+2	68	68
e8m48 + e8m48		11	11	22
e8m48 x e8m48		8	8	16
e8m48 x e8m48 + e8m48		19	19	38
*Assuming e11m48 exponent processing takes 8 ops/input (one complex), 2 ops/output (one complex)				
double = e8m48				
single + double -> double		10		
single x double -> double		6		
double x double -> single		6		
single x single -> double		2		
double + double -> single		9		
divide(e8m48)*	6	70	70	140
recip(e8m48)*	6	56	56	112
sqrt(e8m48)**	8	50	50	100
*Assuming 2 iterations of Newton-Raphson, but probably 3 IRL				
**Assuming e-SQRT and e-RSQRT operations have same throughput				
double = e11m48				
e11m48 + e8m24	20 + 8	16	16	32
e11m48 x e8m24	28 + 8	16	16	32
divide(e11m48)*	6 + 24 + 8	70 + 6 + 16	92	162
recip(e11m48)*	6 + 16 + 4	56 + 4 + 8	68	124
sqrt(e11m48)*	8 + 16 + 4	50 + 4 + 8	62	112
*Too costly for e11m48 ADD, but here we can separate the exponent above 2^2^8, convert to e8m48, and combine later.				
e11 ADD via e8	24 + 8+4	11 + 6 + 16+2	36	47
e11 MUL via e8	24 + 8+4	8 + 6 + 16+2	36	41
e11 FMA via v8	32 + 16+4	19 + 8 + 24+2	53	73
Sources				
https://dl.acm.org/doi/pdf/10.22360/SpringSim.2016.HPC.032				
https://stackoverflow.com/questions/4125033/floating-point-division-vs-floating-point-multiplication				
https://math.stackexchange.com/questions/51150/exact-computational-costs-flop-count-for-algorithms				
https://github.com/danielchalef/openblas-benchmark-m1				
https://github.com/danielchalef/openblas-benchmark-m1/blob/main/dgemm/results/VECLIB/dgemm.veclib.100.txt				
SGEMM GFLOP/s actual				

Time	OpenBLAS 1C	OpenBLAS 8C	Accelerate 1C	Accelerate 2C	Accelerate 8C	GPU e8m24
n=32						
n=64						
n=96						
n=128						
n=160						

DGESVD	0	0	(4/3)n^3	-	-		n=192						
DGEQRF	0	0	(2/3)n^3	-	-		n=256						
DGETRF	0	0	(1/3)n^3	n^2	-		...						
DSYEVD (e-vals)	0	0	(2/3)n^3	-	-		n=2048						
DSYEVD (e-vecs)	0	0	(4/3)n^3	-	-								
ZGEMV	0	0	4n^2	0	0		DGEMM GFLOP/s actual						
ZGEMM	0	0	4n^3	0	0								
ZPOTRF	0	0	(2/3)n^3	(1/2)n^2	-		ZGEMM GFLOP/s actual						
ZGESVD	0	0	(16/3)n^3	-	-								
ZGEQRF	0	0	(8/3)n^3	-	-								
ZGETRF	0	0	(4/3)n^3	n^2	-								
ZSYEVD (e-vals)	0	0	(8/3)n^3	-	-								
ZSYEVD (e-vecs)	0	0	(16/3)n^3	-	-								
FMA GFLOP/s theoretical													
	1-CPU	8-CPU	4-AMX	8-AMX	GPU e8m48	GPU e11m48	i9-13900K	RTX 4090					
Any FP32 workload	103.30	777.22	1652.74	3305.48	10616.00	-	3584.00	82575.36					
Any FP64 workload	51.64	388.60	413.19	826.37	558.74	200.72	1792.00	1290.24					
Time (microseconds)		Accelerate 1C	Accelerate 1C	Accelerate 2C	OpenBLAS 1C	OpenBLAS 2C	GPU e8m48	GPU e11m48					
	n	GFLOPS	Time	Time	Time	Time	Time	Time					
DGEMV	64		2	2									
DGEMM	64		3	3									
DGESVD	64		56	56									
DPOTRF	64		16	16									
DSYEVD (e-vecs)	64		148	148									
CGEMV	64												
CGEMM	64												
CGESVD	64												
CPOTRF	64												
CSYEVD (e-vecs)	64												
DPOTRF	96												
DSYEVD (e-vecs)	96												
CPOTRF	96												
CSYEVD (e-vecs)	96												
DPOTRF	128												
DSYEVD (e-vecs)	128												
CPOTRF	128												
CSYEVD (e-vecs)	128												
DPOTRF	256												
DSYEVD (e-vecs)	256												
CPOTRF	256												
CSYEVD (e-vecs)	256												
DPOTRF	384												
DSYEVD (e-vecs)	384												
CPOTRF	384												
CSYEVD (e-vecs)	384												
DPOTRF	512												

DSYEVD (e-vecs)	512												
CPOTRF	512												
CSYEVD (e-vecs)	512												
DPOTRF	768												
DSYEVD (e-vecs)	768												
CPOTRF	768												
CSYEVD (e-vecs)	768												
GFLOPS		Accelerate MC	OpenBLAS MC	GPU e8m48	GPU e11m48	Accelerate MC	OpenBLAS MC						
	n	GFLOPS	GFLOPS	GFLOPS	GFLOPS	Time (us)	Time (us)						
DGEMM	32	93	21	2	2	1							
ZGEMM	32	54	9	2	2	8							
DGEMM	48	142	36	7	7	2							
ZGEMM	48	63	36	7	7	16							
DGEMM	64	177	40	18	18	3							
ZGEMM	64	83	67	18	18	24							
DGEMM	96	227	61	52	52	8							
ZGEMM	96	122	75	52	52	60							
DGEMM	128	247	79	235	200	17							
ZGEMM	128	158	29	235	200	104							
DGEMM	192	290	48	406	200	48							
ZGEMM	192	159	86	406	200	356							
DGEMM	256	333		558	200	101							
ZGEMM	256			558	200								
DGEMM	384	622		558	200	184							
ZGEMM	384			558	200								
DGEMM	512	616		558	200	386							
ZGEMM	512			558	200								
DGEMM	768	696		558	200	1302							
ZGEMM	768			558	200								
DGEMM	1024	442		558	200	4859							
ZGEMM	1024			558	200								
DGEMM	1536	566		558	200	12805							
ZGEMM	1536			558	200								
NumPy Eigh (ZHEEVD)	Accelerate	Accelerate	OpenBLAS	OpenBLAS									
Size	Time (us)	GFLOPS	Time (us)	GFLOPS									
32	38	4.6	32	5.5									
48	79	7.5	72	8.2									

64	104	13.4	95	14.7									
96	285	16.6	260	18.1									
128	468	23.9	478	23.4									
192	1376	27.4	1273	29.7									
256	3218	27.8	2688	33.3									
384	9893	30.5	8245	36.6									
512	22507	31.8	19194	37.3									
768	66005	36.6	55982	43.2									
1024	156395	36.6	103882	55.1									
1536	437552	44.2	238301	81.1									
2048	1339815	34.2	595944	76.9									
NumPy Eig (DSYEVD)	Accelerate	Accelerate	OpenBLAS	OpenBLAS									
Size	Time (us)	GFLOPS	Time (us)	GFLOPS									
32	27	1.6	22	2.0									
48	53	2.8	45	3.3									
64	79	4.4	76	4.6									
96	175	6.7	171	6.9									
128	292	9.6	267	10.5									
192	526	17.9	580	16.3									
256	852	26.3	1028	21.8									
384	1938	39.0	2676	28.2									
512	4161	43.0	5528	32.4									
768	9819	61.5	14381	42.0									
1024	24424	58.6	31705	45.2									
1536	61833	78.1	81547	59.3									
2048	167896	68.2	150412	76.1									
3072			456313	84.7									