All measurements are in GFLOPS. The CPU sometimes uses AMX, and multicore only harms CPU performance.					
N x N x N	CPU F64	CPU F32	GPU F32	GPU F32 x 4 (duplicate)	GPU F32 x 4 (unique)
8	2	2	0	0	0
16	12	12	0	0	0
24	31	34	2	1	1
32	78	98	6	4	4
48	143	204	18	15	14
64	209	434	49	36	36
96	267	745	85	103	103
128	304	940	264	271	267
192	323	1117	398	813	821
256	333	1223	1266	1792	1790
384	622	1303	3040	3710	3667
512	616	2282	5293	6804	6770
768	696	2679	6058	7297	7292
1024	442	2262	8048	8306	8263
1536	566	2021	8203	8389	8376
2048	536	1978	8362	8417	8370
3072	516	2058	8351	8368	8289
4096	520	1957	8162	8171	8136
6144	504	1957	8126	8050	8034
8192	524	1998	7898	7209	7426
10240	508	1965	7688	6937	6959
12288	522	1989	7440	6912	6699
Fastest recorded trial:			Theoretical Max		
simdgroup_matrix<F16>		8138	10616	*2816x2816 x 4 (duplicate)	
simdgroup_matrix<F32>		8435	10616	*2176x2176 x 4 (duplicate)	
AMX (CPU)	F16	2392	3305	*4608x4608	
AMX (CPU)	F32	2746	3305	*960x960	
AMX (CPU)	F64	700	826	*672x672	