Figure: Scatter plot with x-axis "Exe. Time Needed to Reach Peak Mem. Consumption (s)" and y-axis "Peak Memory Consumption" (GB).

Legend:
- llama-70B-BS8-L4096-stage0
- llama-8B-BS16-L4096-stage0
- granite-3B-BS32-L1024-stage0
- granite-8B-BS16-L1024-stage0
- mistral-7B-BS16-L1024-stage0
- gpt2-40B-BS16-L1024-stage0
- 16 GB/s (PCIe Gen3)
- 32 GB/s (PCIe Gen4)