

# Overlapping Computation and Communication for Advection on Hybrid Parallel Computers

James B White III (Trey)

trey@ucar.edu

National Center for Atmospheric Research

Jack Dongarra

dongarra@eecs.utk.edu

University of Tennessee, Knoxville

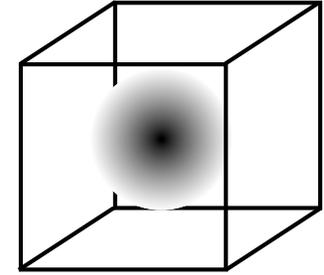
Programming Weather, Climate, and Earth-System Models  
on Heterogeneous Multi-Core Platforms

NCAR, September 8, 2011

based on work first presented at IPDPS, Anchorage, AK, May 17, 2011

*Portions of this work were funded by the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research, both of the US Department of Energy. This research used resources of the OLCF at Oak Ridge National Laboratory and of NERSC at Lawrence Berkeley National Laboratory, both of which are supported by the Office of Science of the US Department of Energy.*

# Test Case



- Linear advection with constant uniform velocity
- Three-dimensional cube with periodic boundaries
- Advect Gaussian wave through cube corner back to original position
- Strong scaling,  $420 \times 420 \times 420$
- Explicit 2nd-order single-stage integration,  $3 \times 3 \times 3$  centered stencil, 64-bit precision

# Computers

System	JaguarPF	Hopper II	Lens	Yona
Compute nodes	18688	6392	31	16
Memory per node (GB)	16	32	64	32
AMD Opteron sockets per node	2	2	4	2
Cores per Opteron socket	6	12	4	6
Opteron clock (GHz)	2.6	2.1	2.3	2.6
Interconnect	Cray SeaStar 2+	Cray Gemini	DDR Infiniband	QDR Infiniband
MPI	Cray MPT 4.0.0	Cray MPT 5.1.3	OpenMPI 1.3.3	OpenMPI 1.7a1
NVIDIA Tesla GPU	–	–	C1060	C2050
GPU memory (GB)	–	–	4	3

# Computers

System	JaguarPF	Hopper II	Lens	Yona
Compute nodes	18688	6392	31	16
Memory per node (GB)	16	32	64	32
AMD Opteron sockets per node	2	2	4	2
Cores per Opteron socket	6	12	4	6
Opteron clock (GHz)	2.6	2.1	2.3	2.6
Interconnect	Cray SeaStar 2+	Cray Gemini	DDR Infiniband	QDR Infiniband
MPI	Cray MPT 4.0.0	Cray MPT 5.1.3	OpenMPI 1.3.3	OpenMPI 1.7a1
NVIDIA Tesla GPU	–	–	C1060	C2050
GPU memory (GB)	–	–	4	3

# Computers

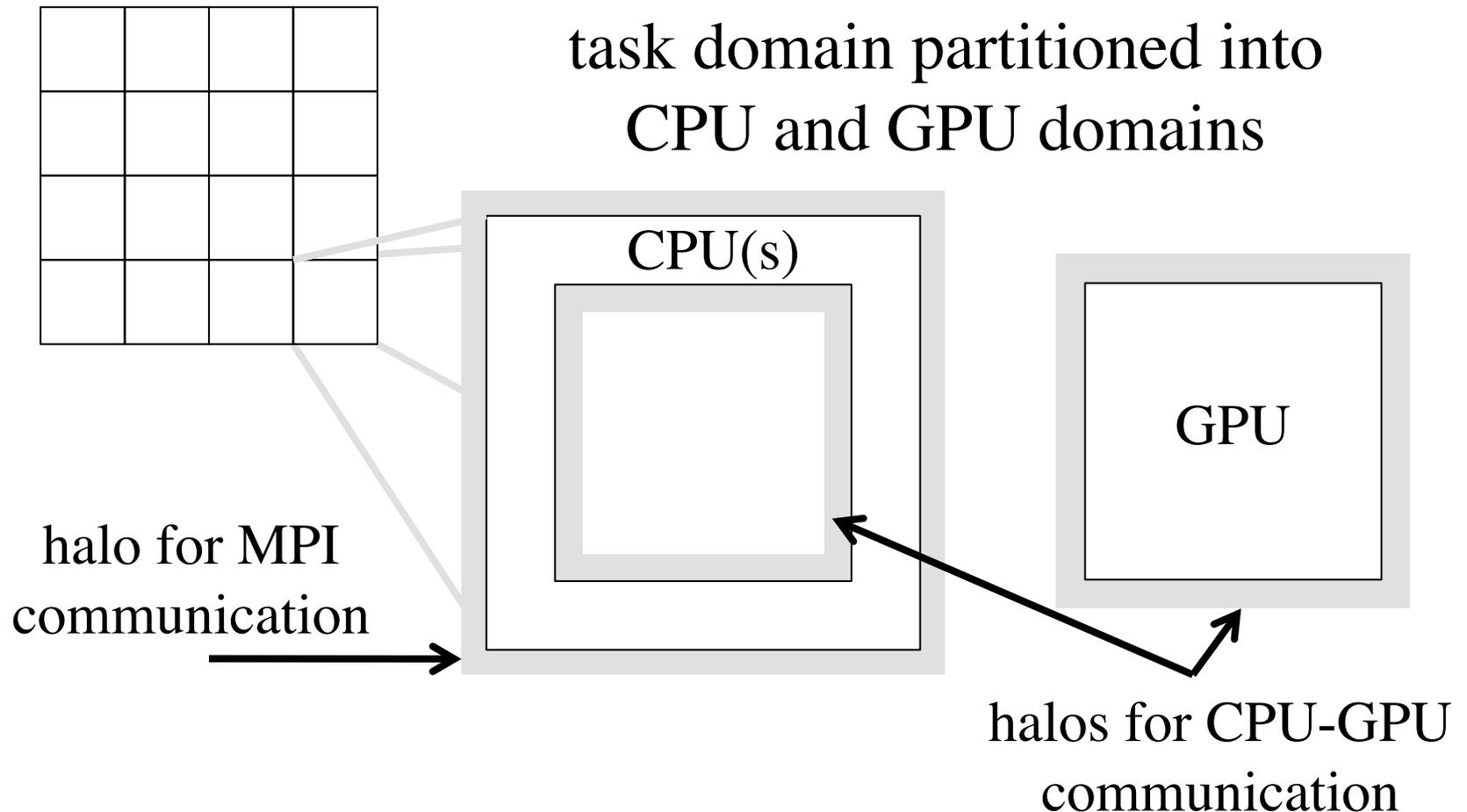
System	JaguarPF	Hopper II	Lens	Yona
Compute nodes	18688	6392	31	16
Memory per node (GB)	16	32	64	32
AMD Opteron sockets per node	2	2	4	2
Cores per Opteron socket	6	12	4	6
Opteron clock (GHz)	2.6	2.1	2.3	2.6
Interconnect	Cray SeaStar 2+	Cray Gemini	DDR Infiniband	QDR Infiniband
MPI	Cray MPT 4.0.0	Cray MPT 5.1.3	OpenMPI 1.3.3	OpenMPI 1.7a1
NVIDIA Tesla GPU	–	–	C1060	C2050
GPU memory (GB)	–	–	4	3

# Implementations

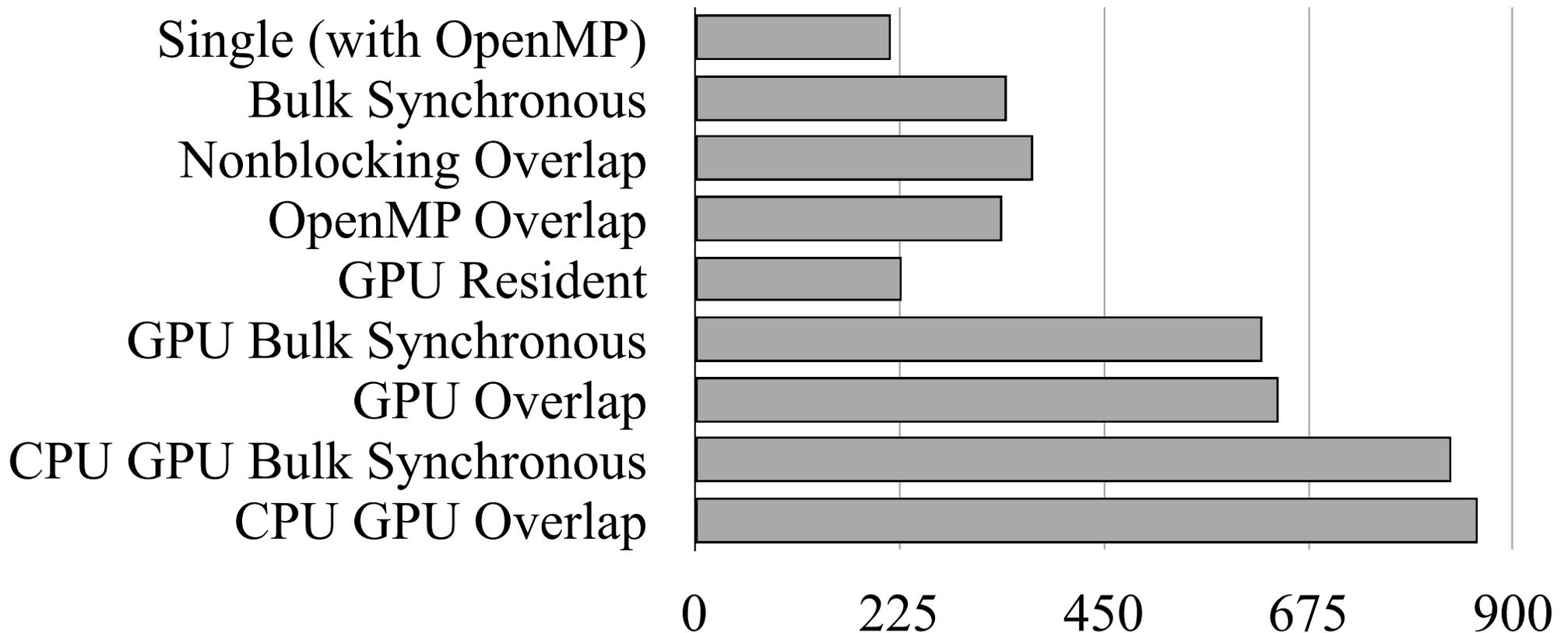
- Single task (Fortran + OpenMP)
- Bulk-synchronous MPI
- MPI using nonblocking communication for overlap
- MPI using OpenMP threading for overlap
- GPU resident (CUDA Fortran)
- GPU with bulk-synchronous MPI
- GPU with MPI overlap using CUDA streams
- CPU and GPU computation with bulk-synchronous MPI
- CPU and GPU computation partitioned for overlap with nonblocking MPI and CPU-GPU communication

# CPU-GPU Domain Decomposition

global domain decomposed  
into MPI-task domains

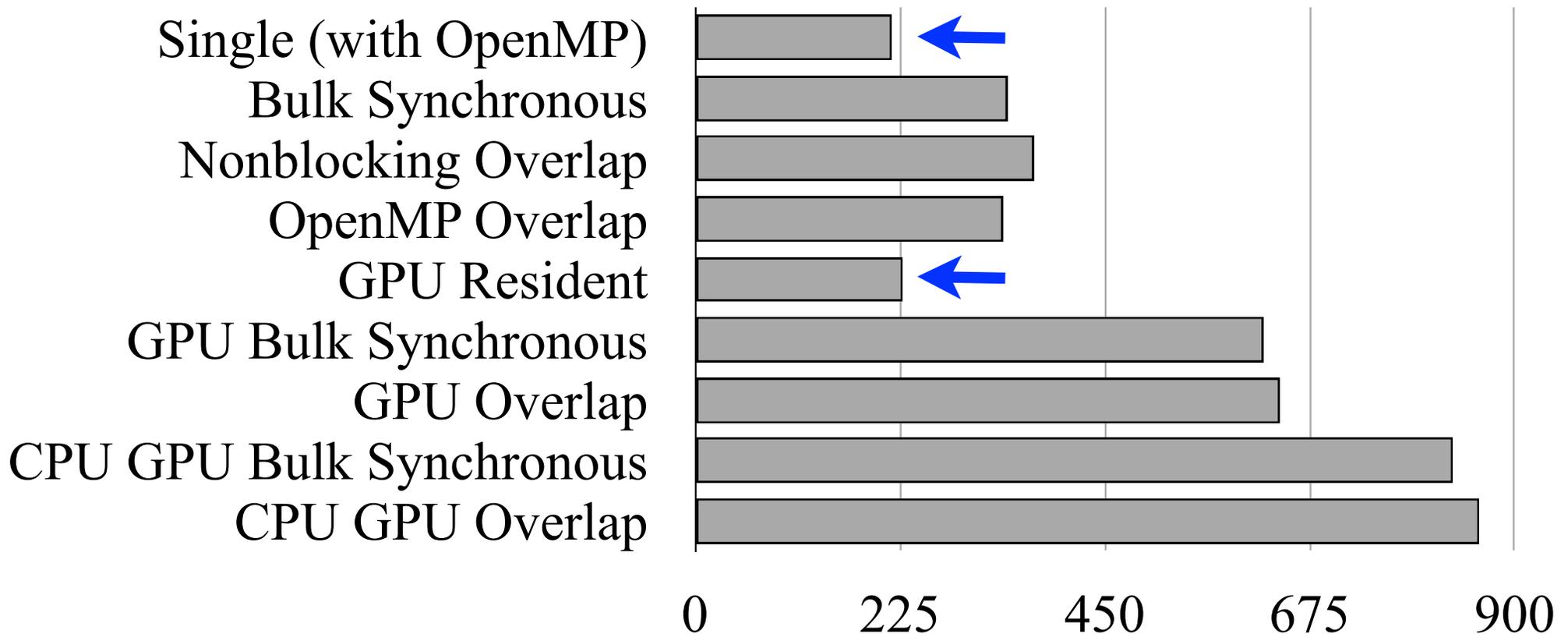


# Lines of Code



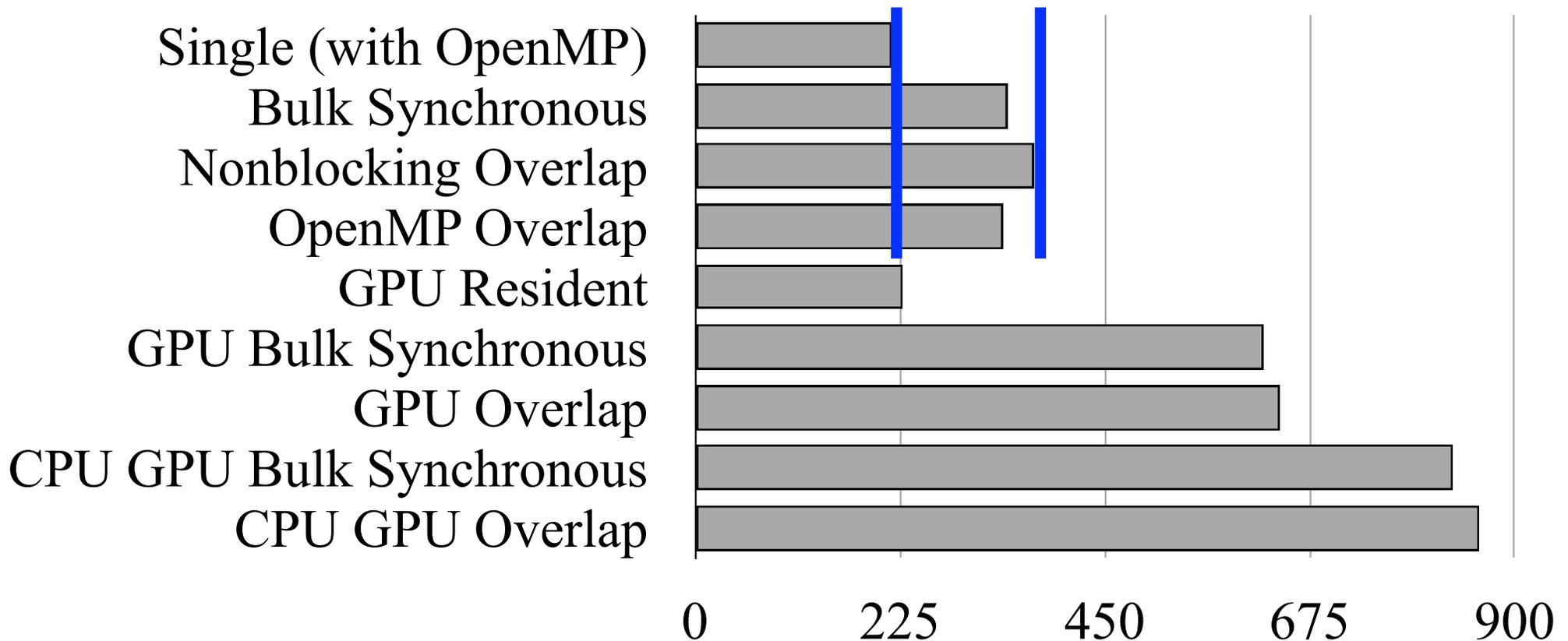
# Lines of Code

*Similar*



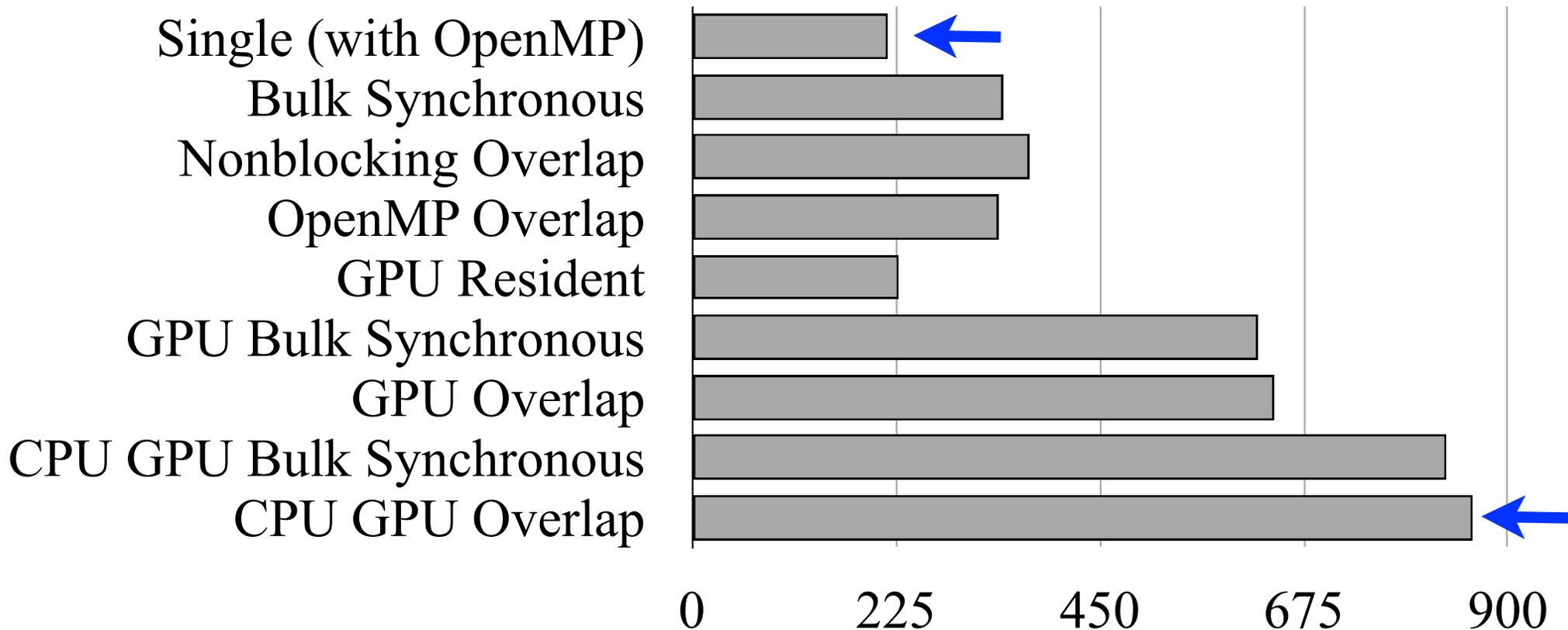
# Lines of Code

*MPI adds 50-75%*

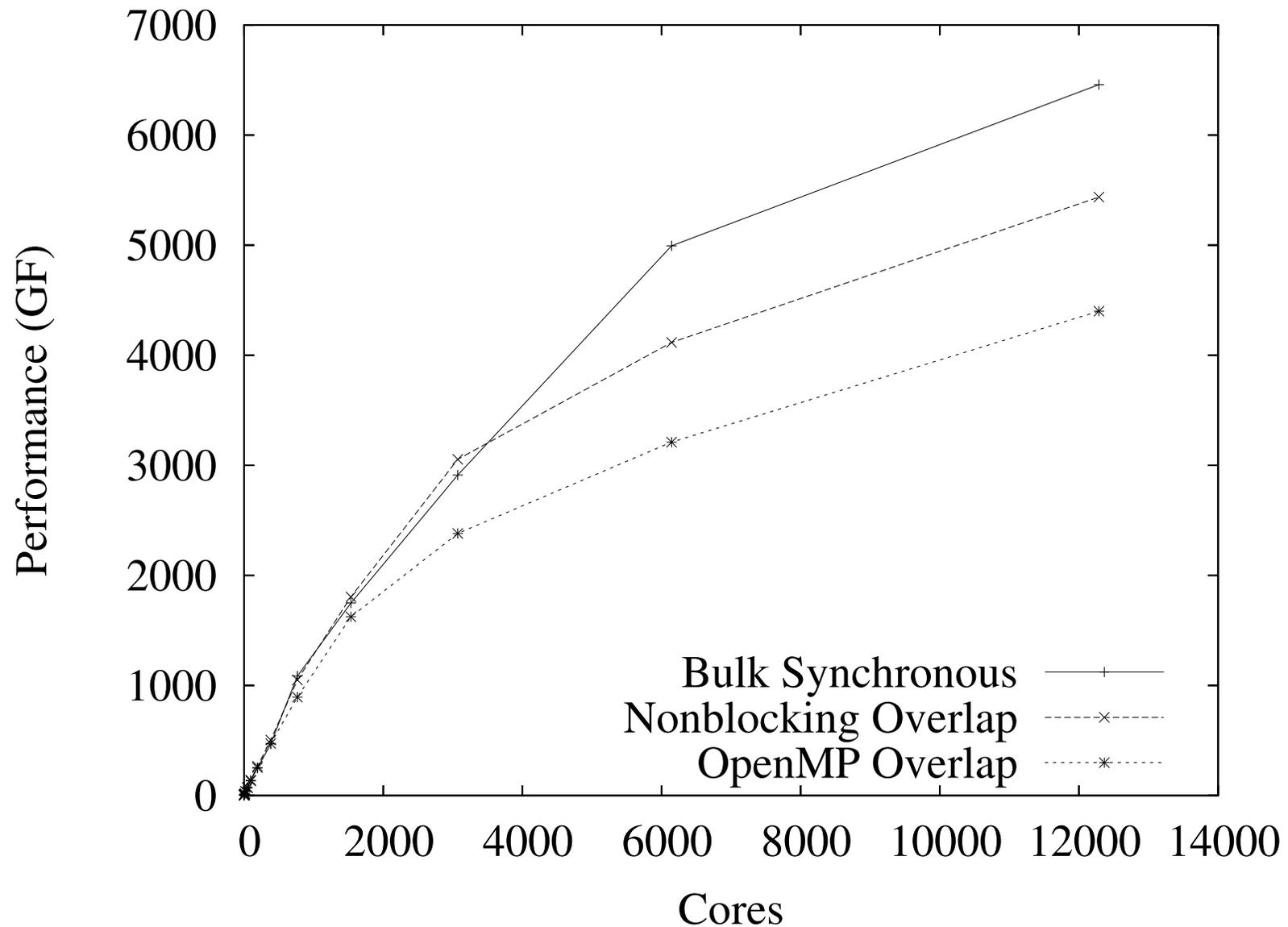


# Lines of Code

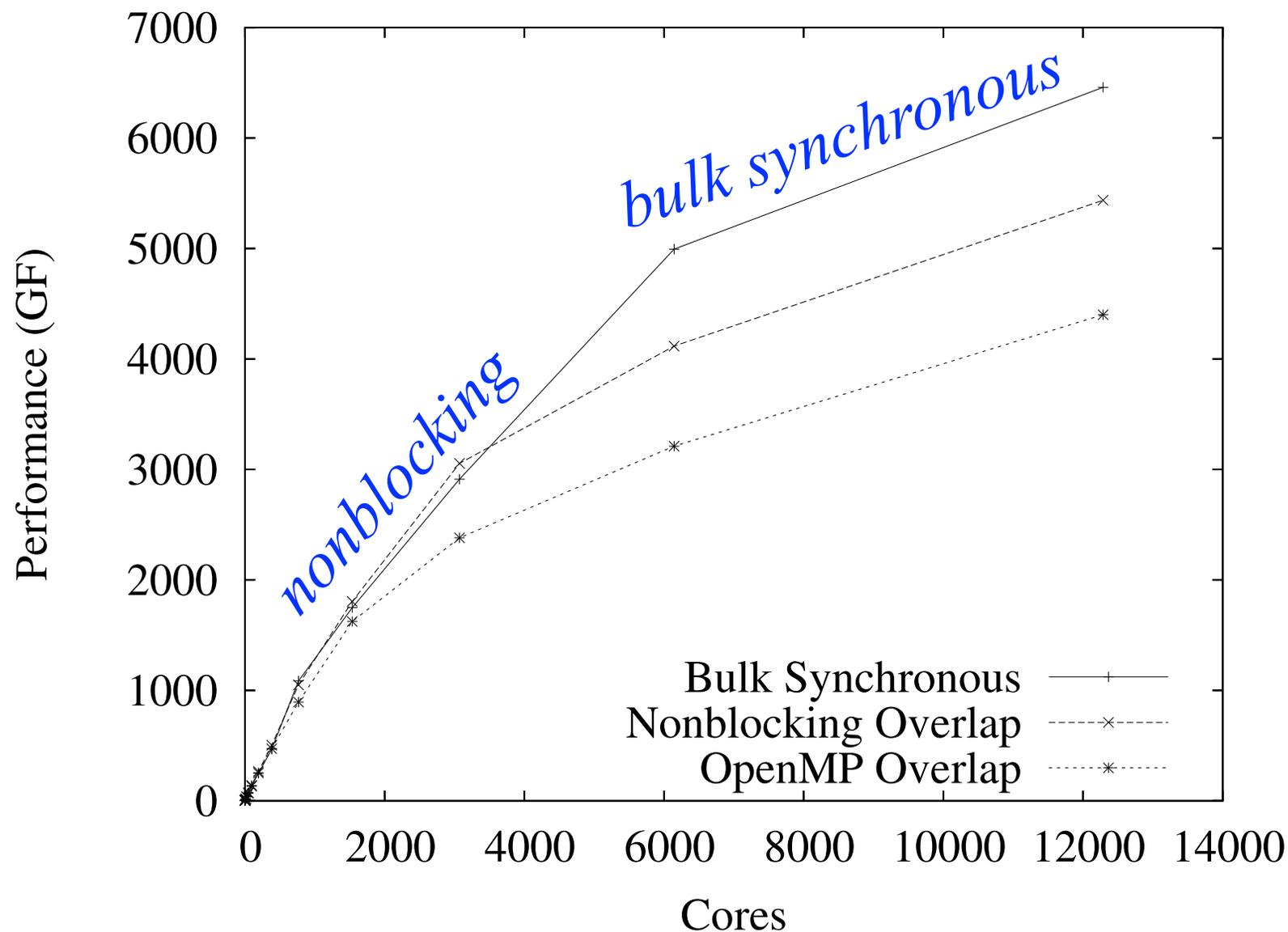
*4 times the code*



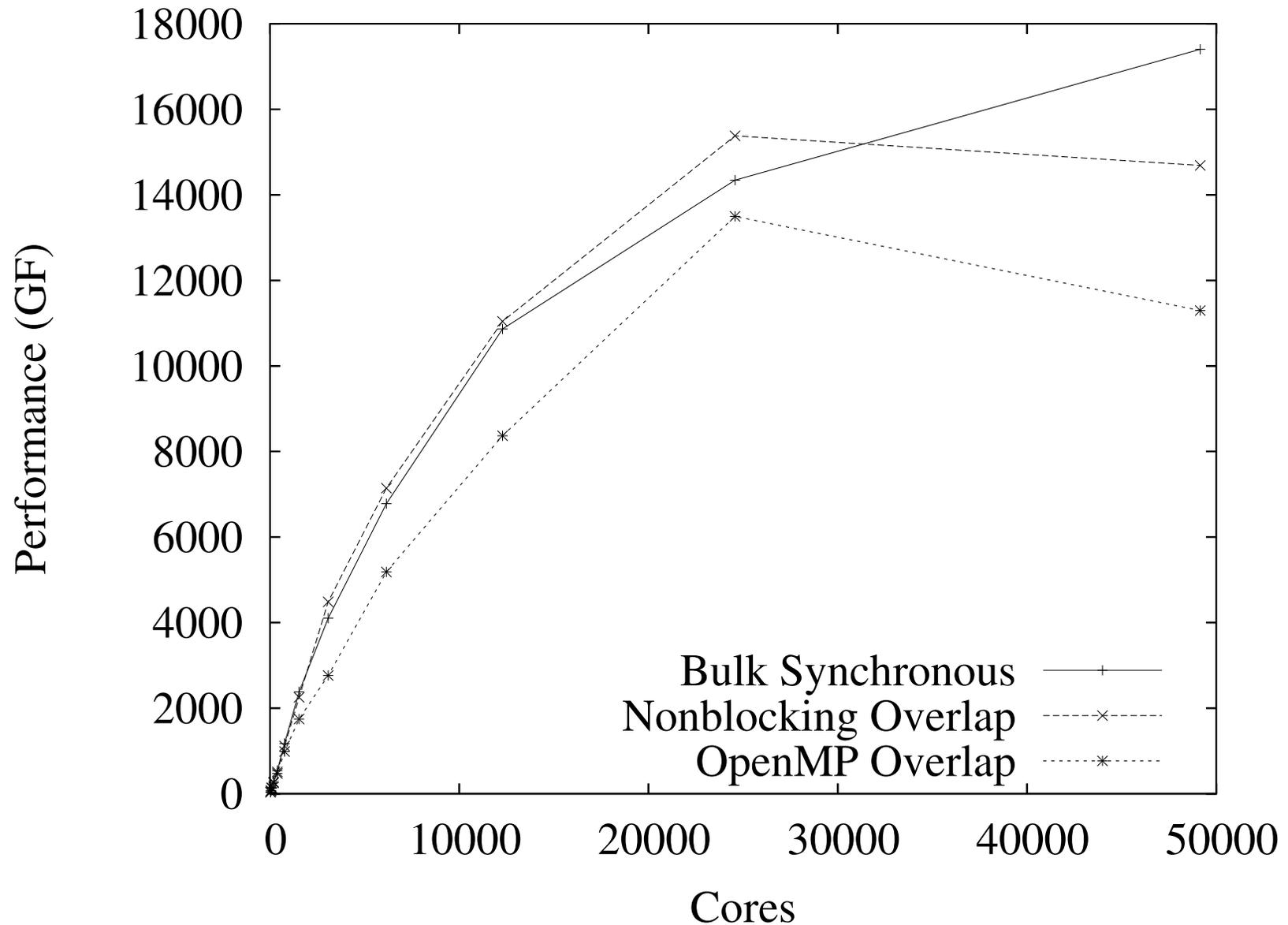
# Best JaguarPF Performance



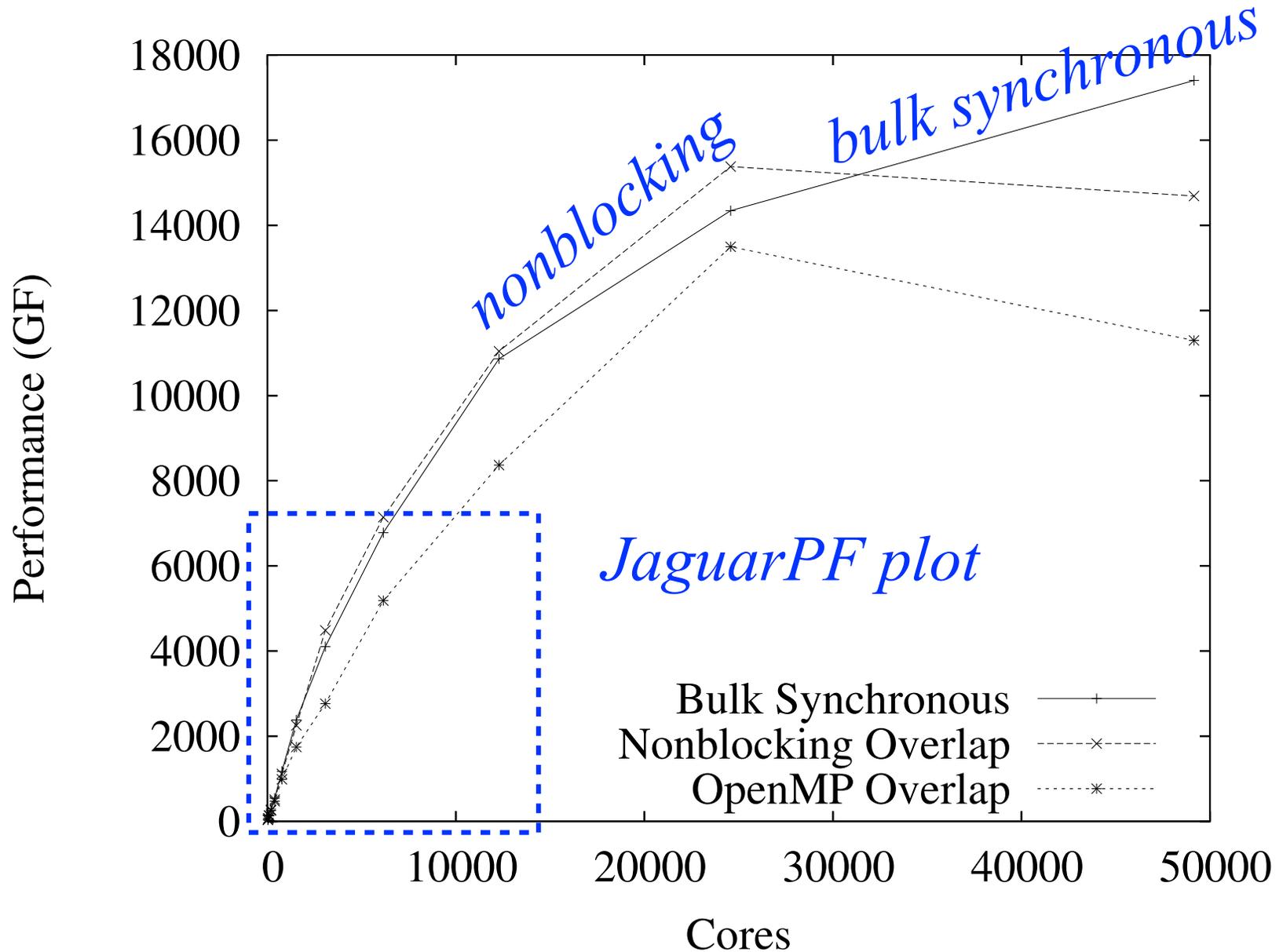
# Best JaguarPF Performance



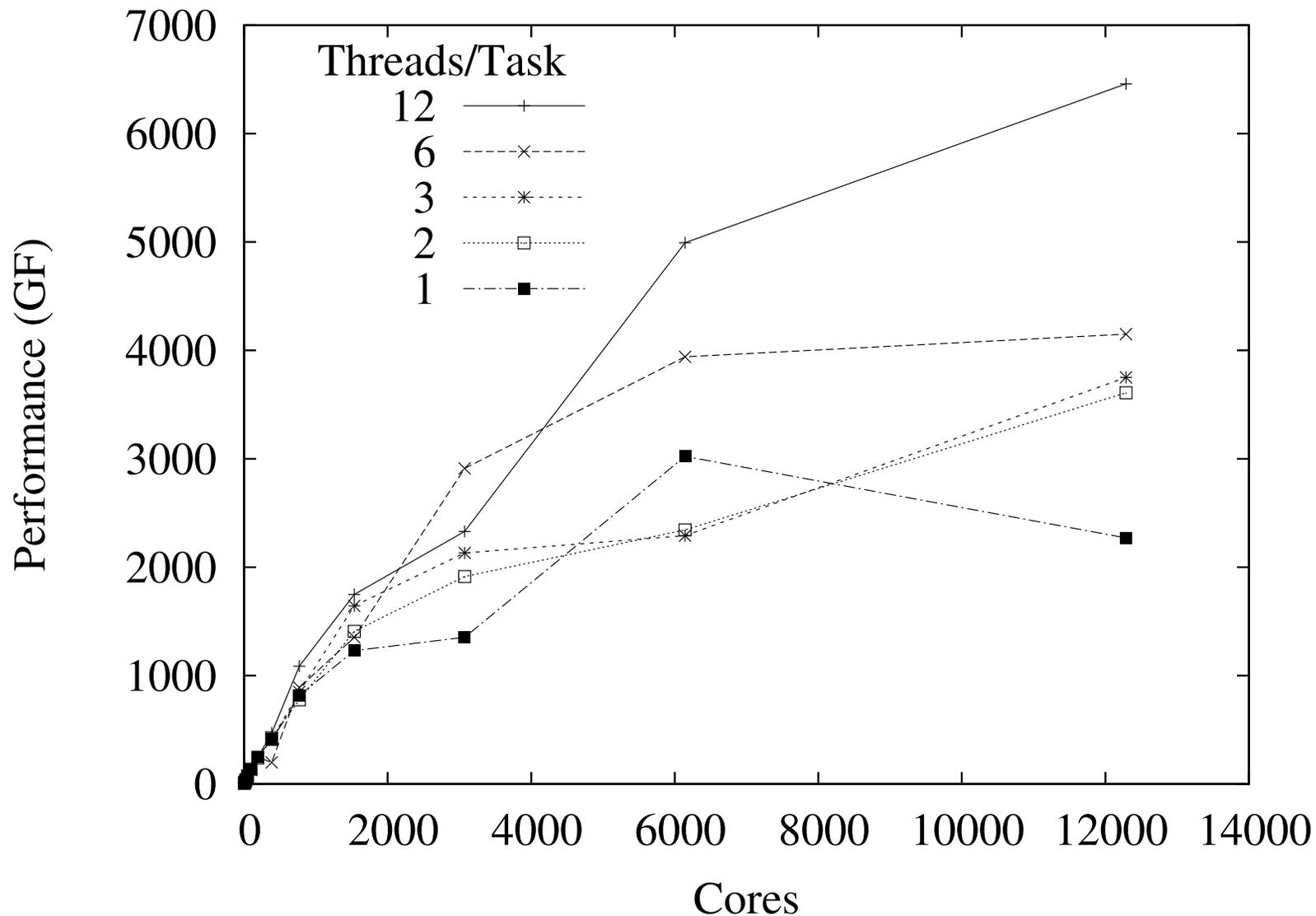
# Best Hopper-II Performance



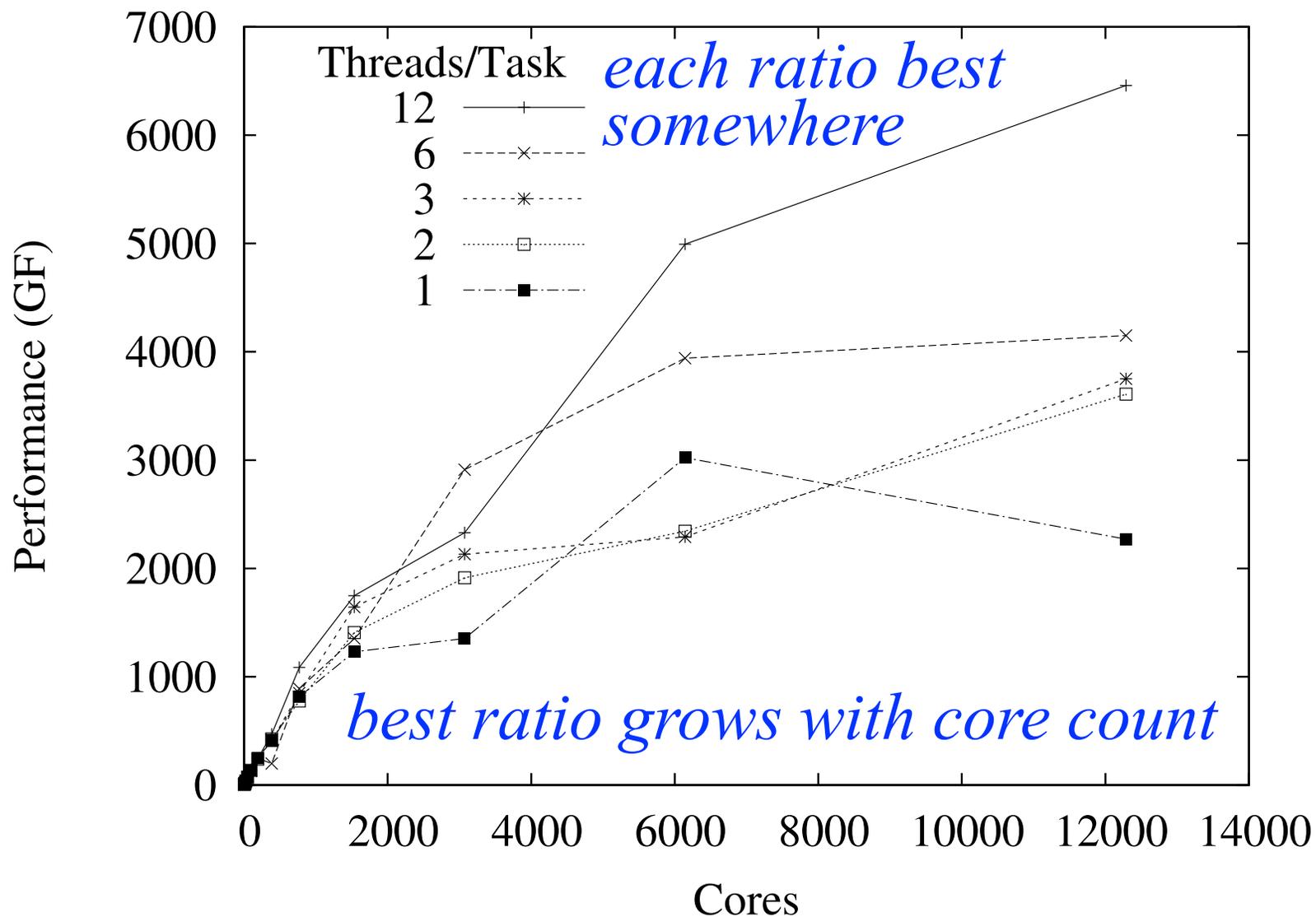
# Best Hopper-II Performance



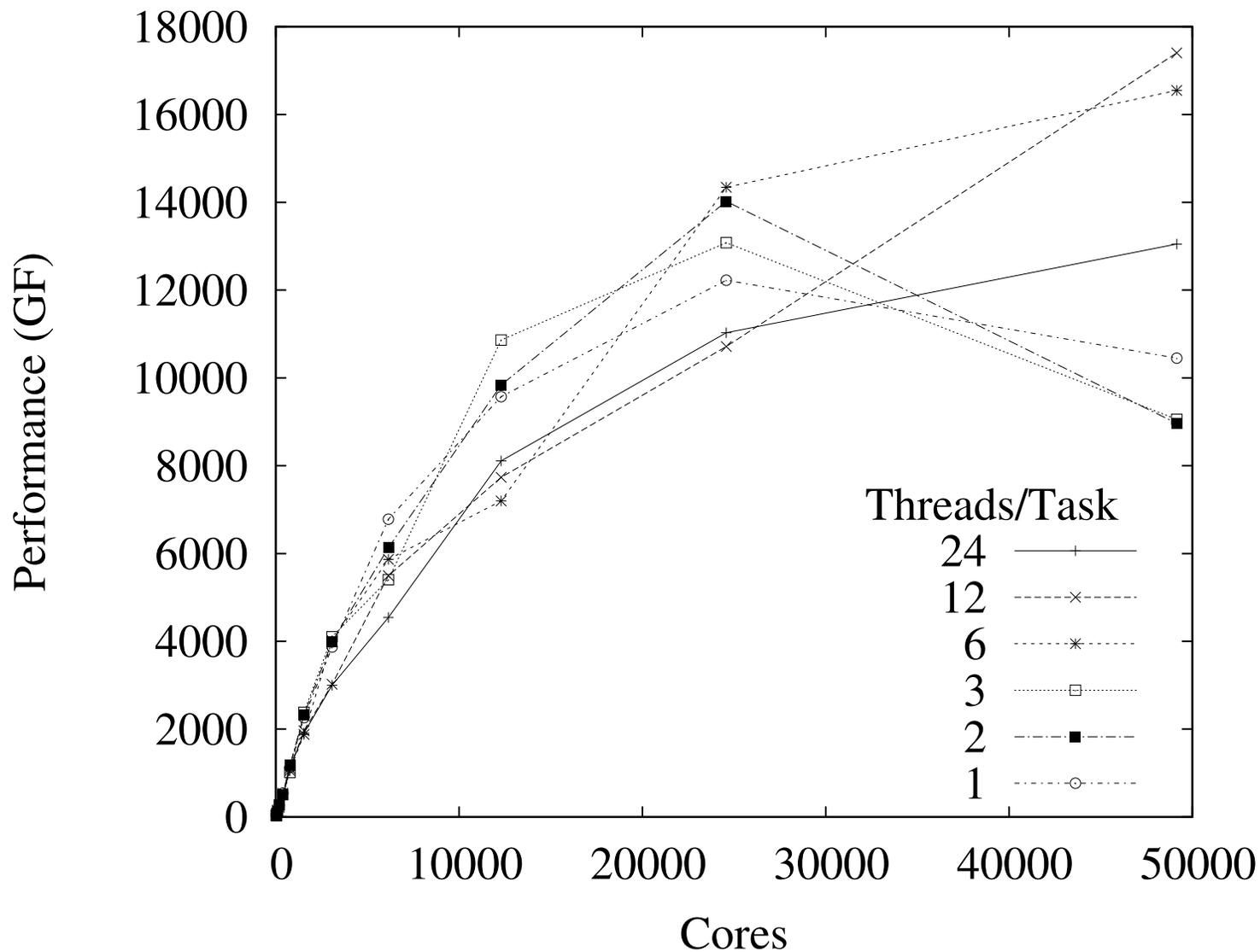
# Bulk-Synchronous Performance on JaguarPF



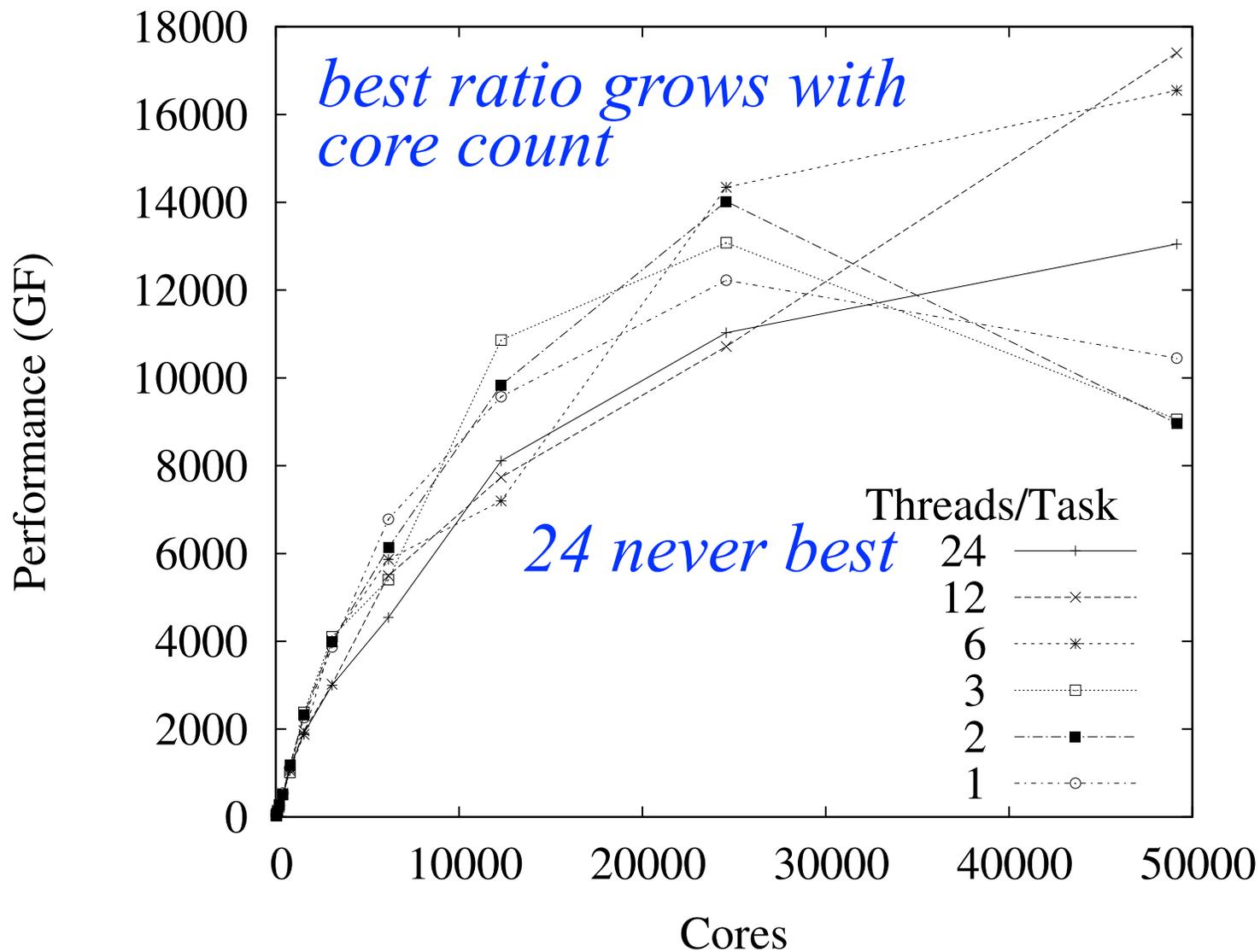
# Bulk-Synchronous Performance on JaguarPF



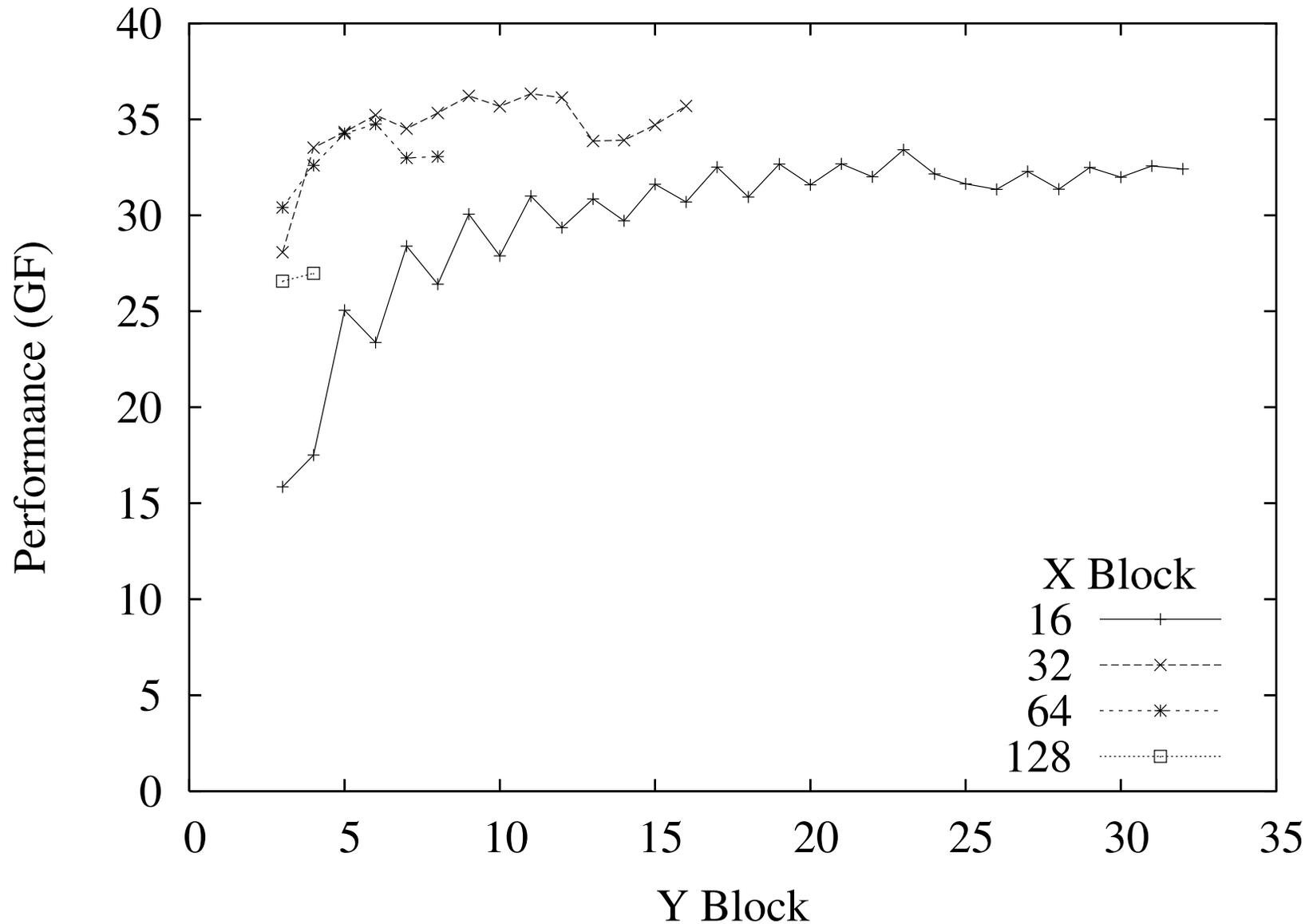
# Bulk-Synchronous Performance on Hopper II



# Bulk-Synchronous Performance on Hopper II



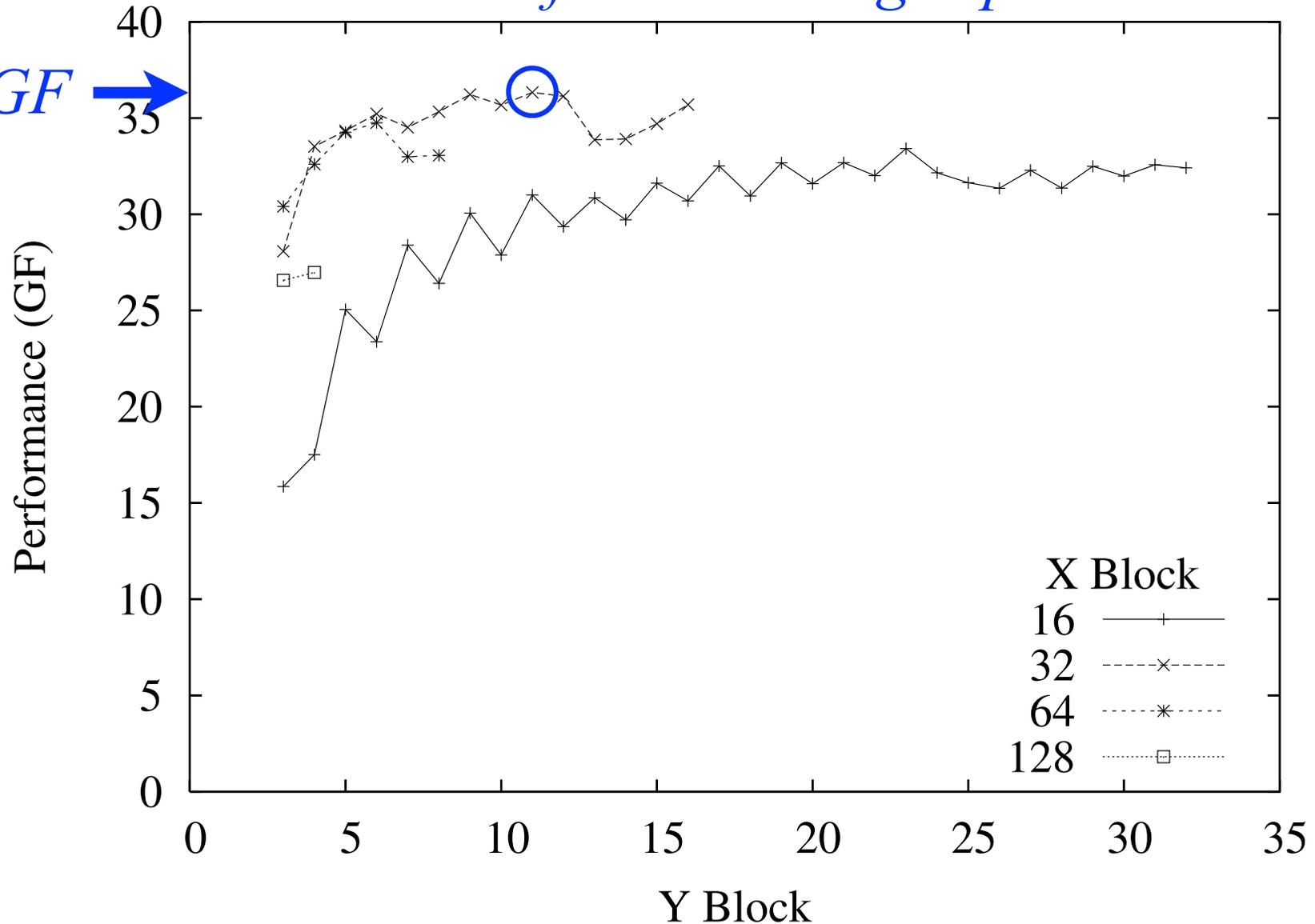
# GPU-Resident Performance on Lens



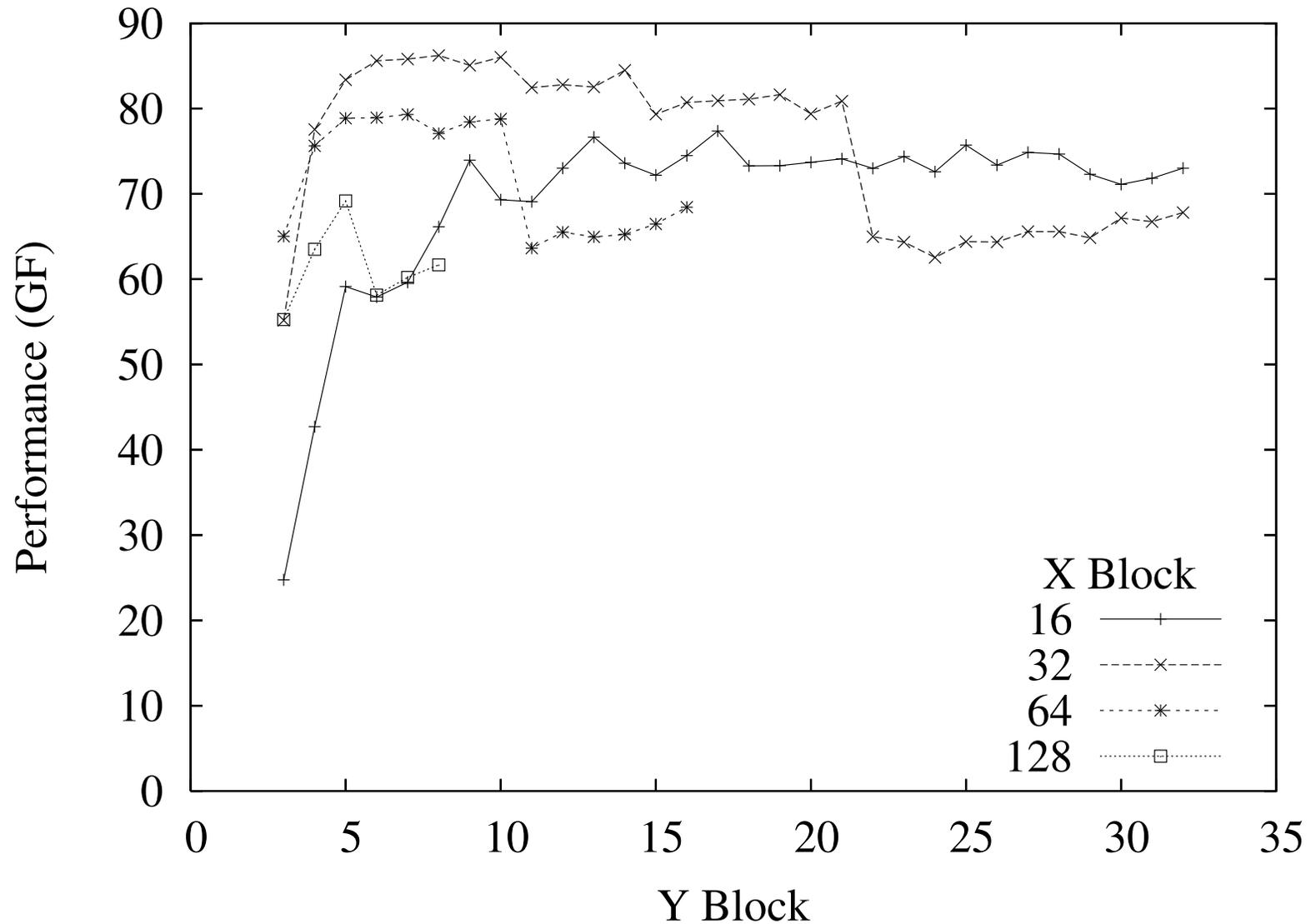
# GPU-Resident Performance on Lens

*32x11 used for remaining experiments*

*36.3 GF* →



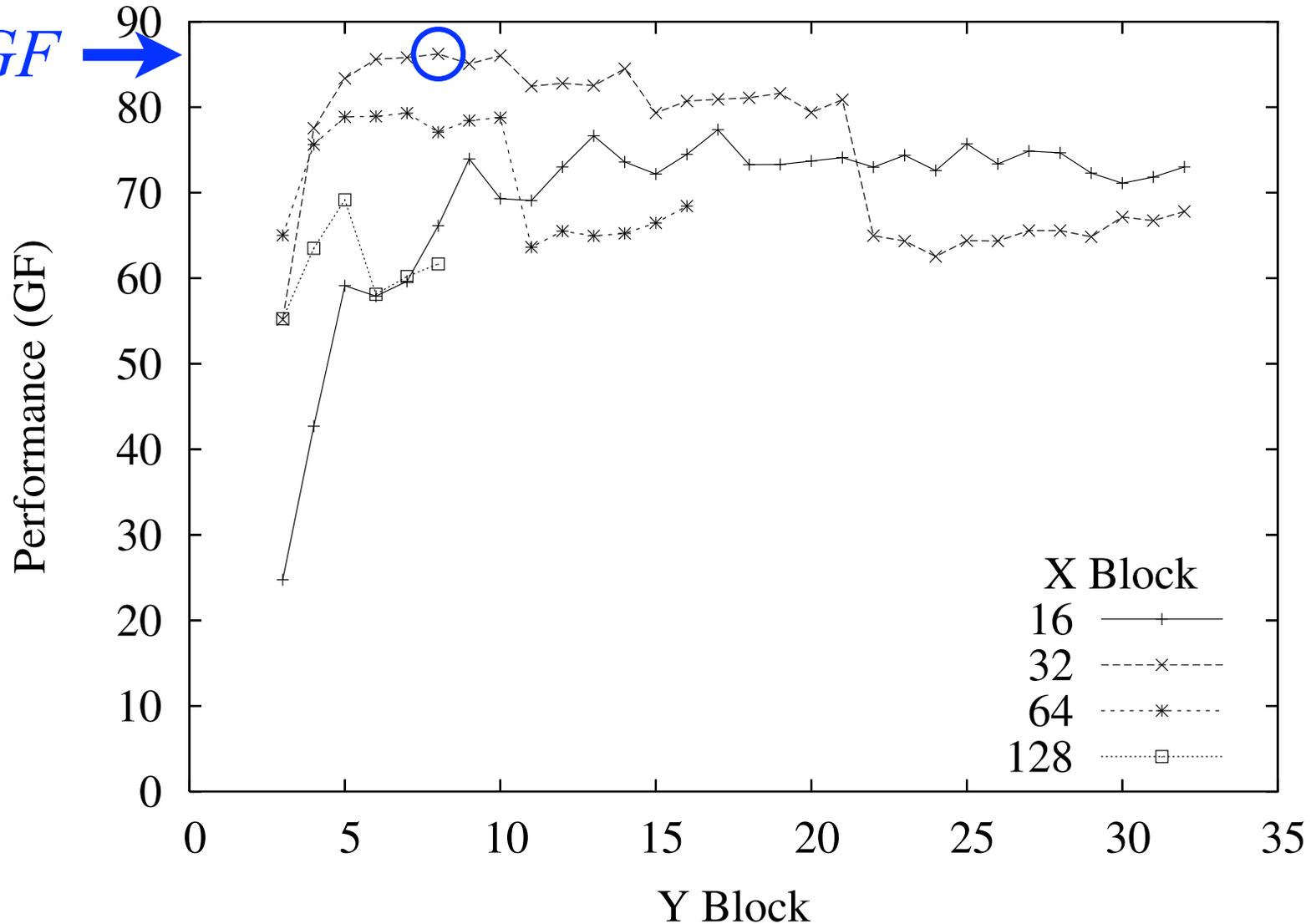
# GPU-Resident Performance on Yona



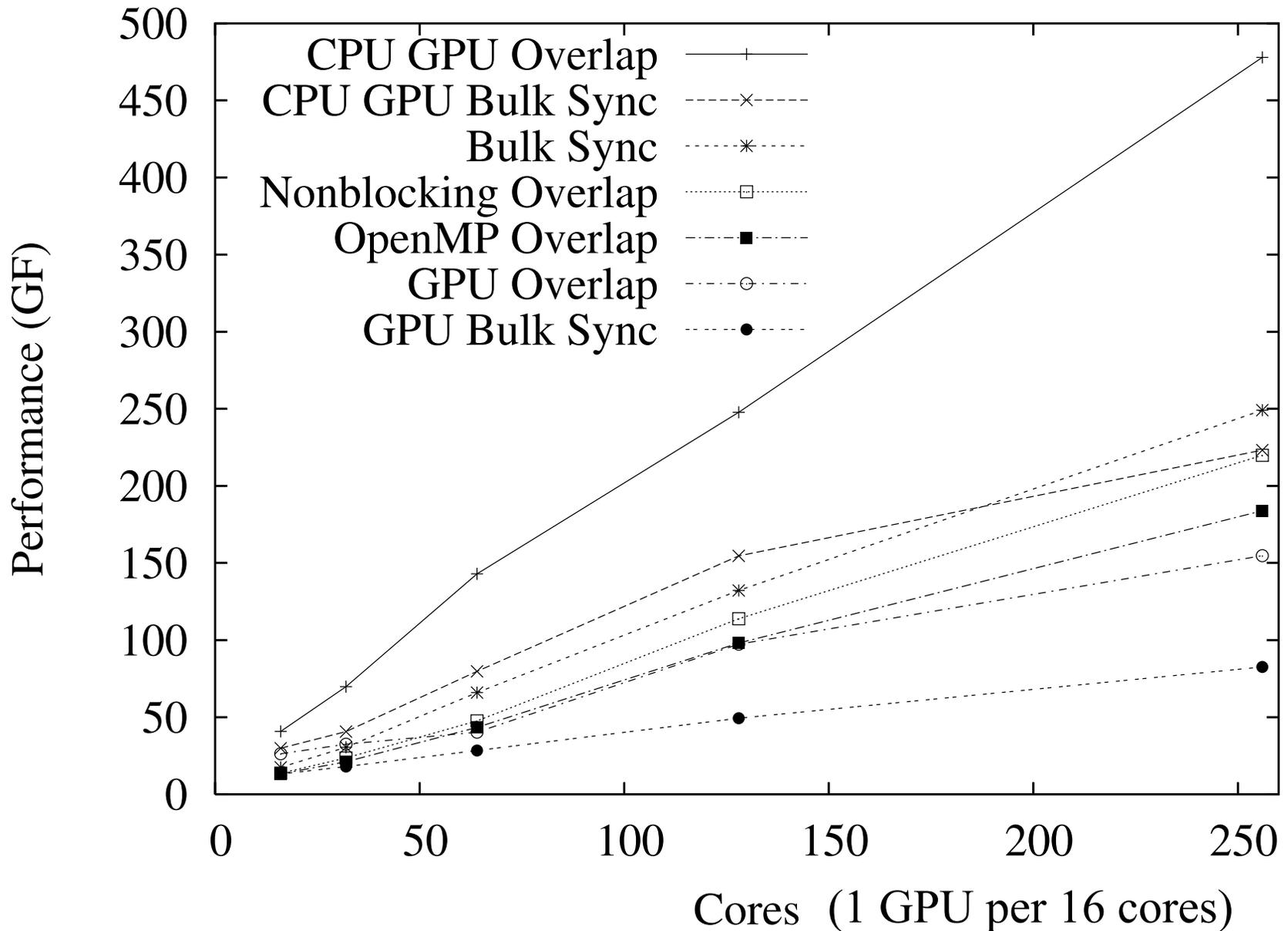
# GPU-Resident Performance on Yona

*32x8 used for remaining experiments*

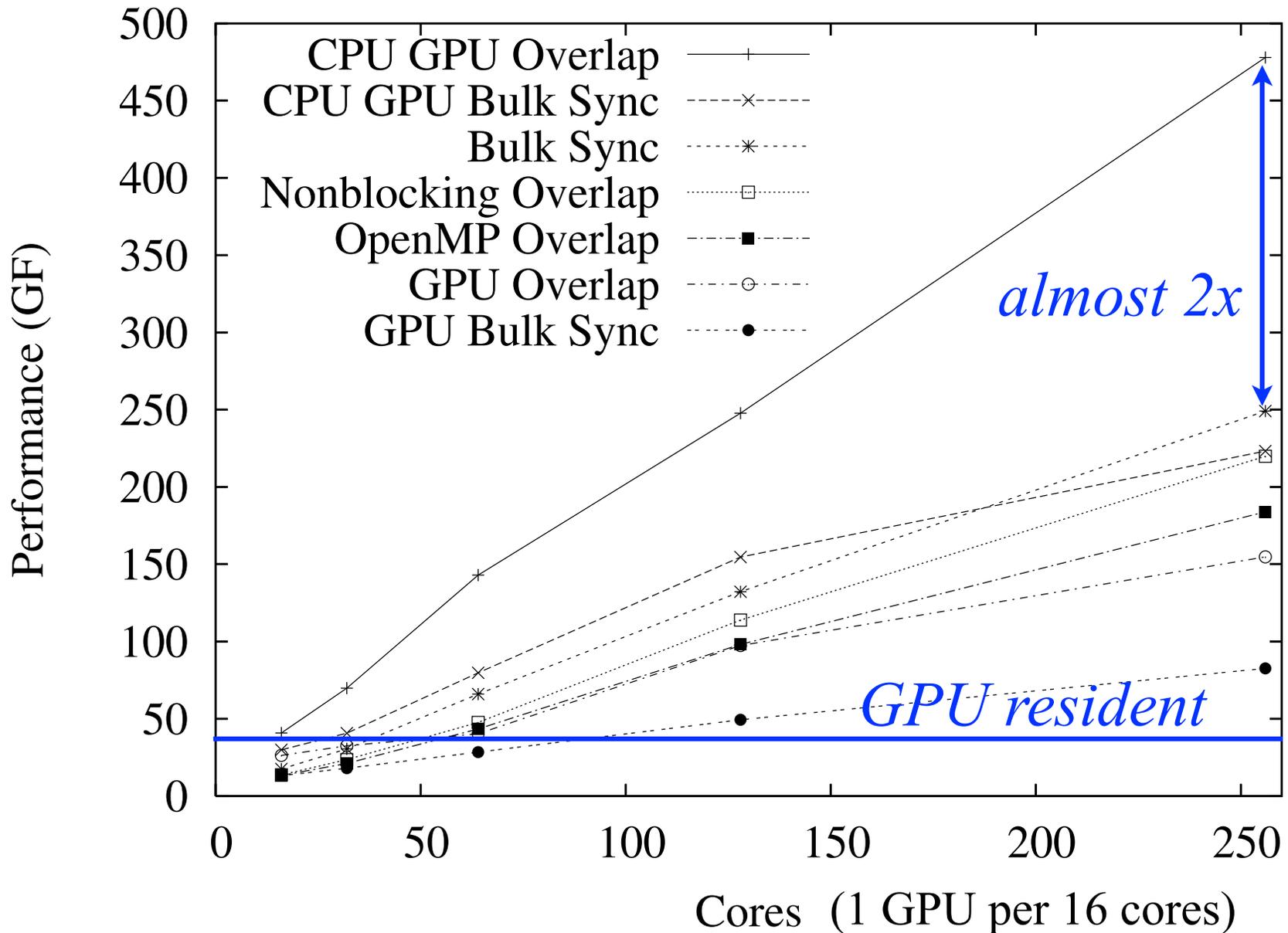
*86.2 GF* →



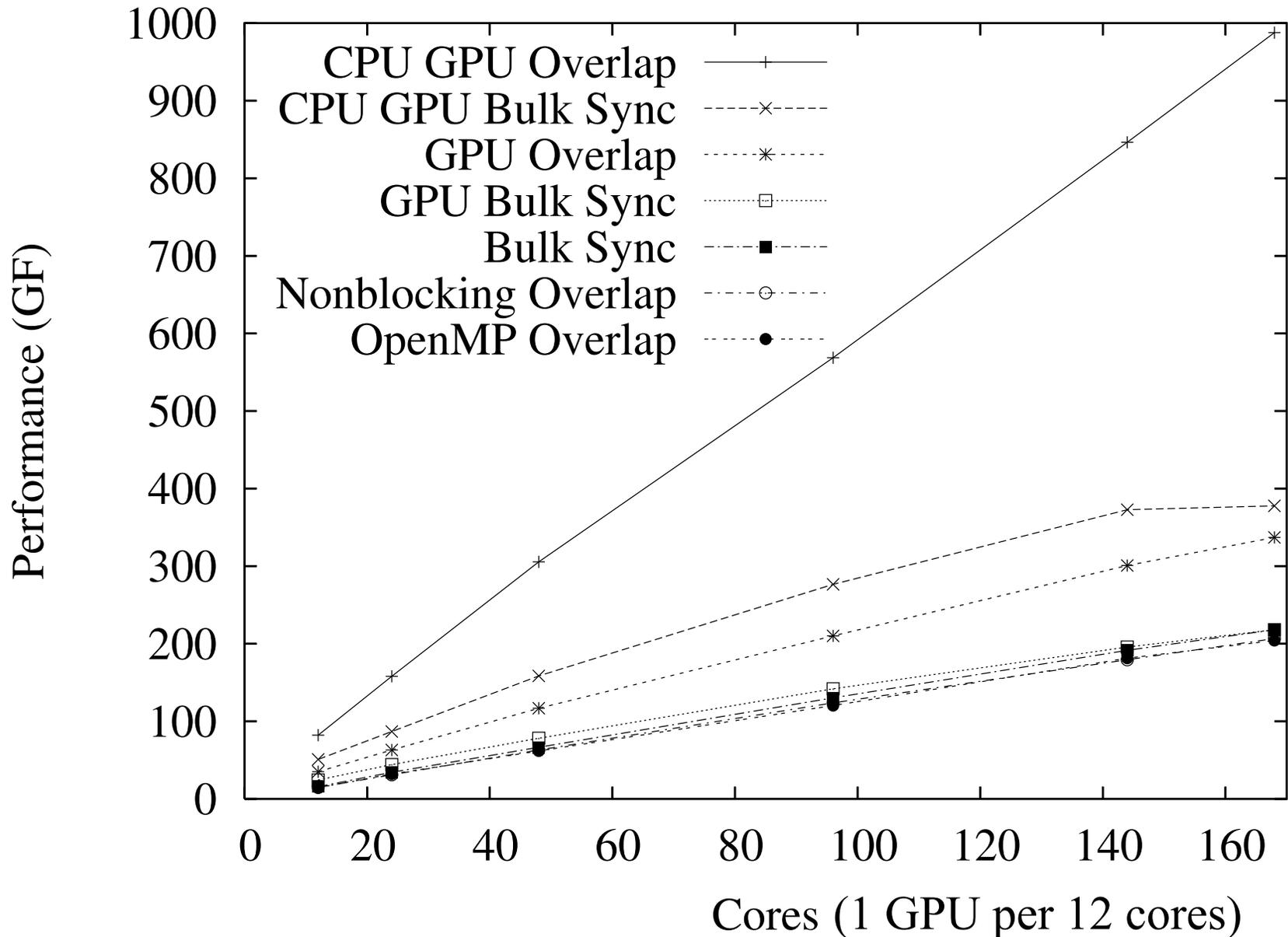
# Best Performance on Lens



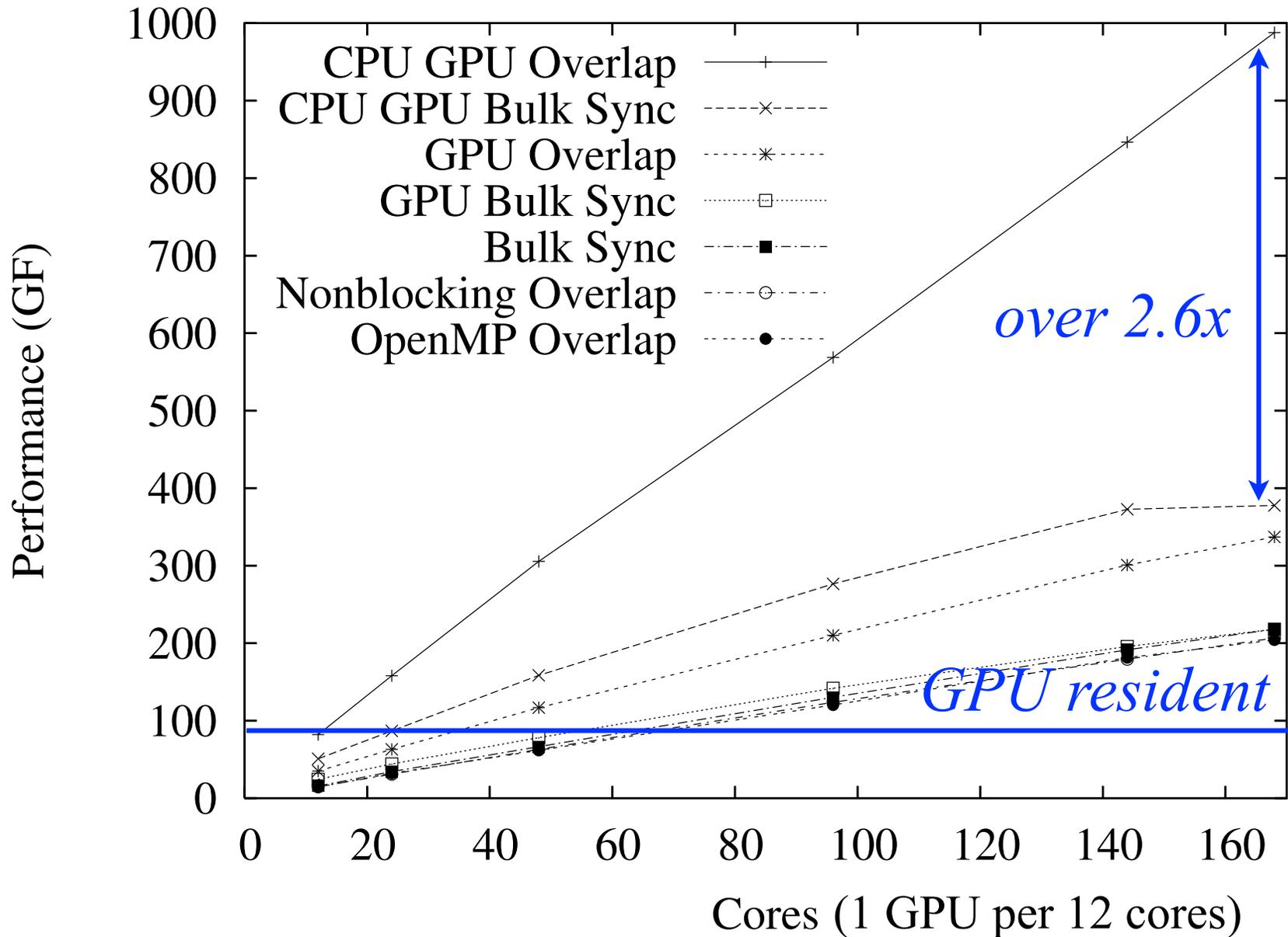
# Best Performance on Lens



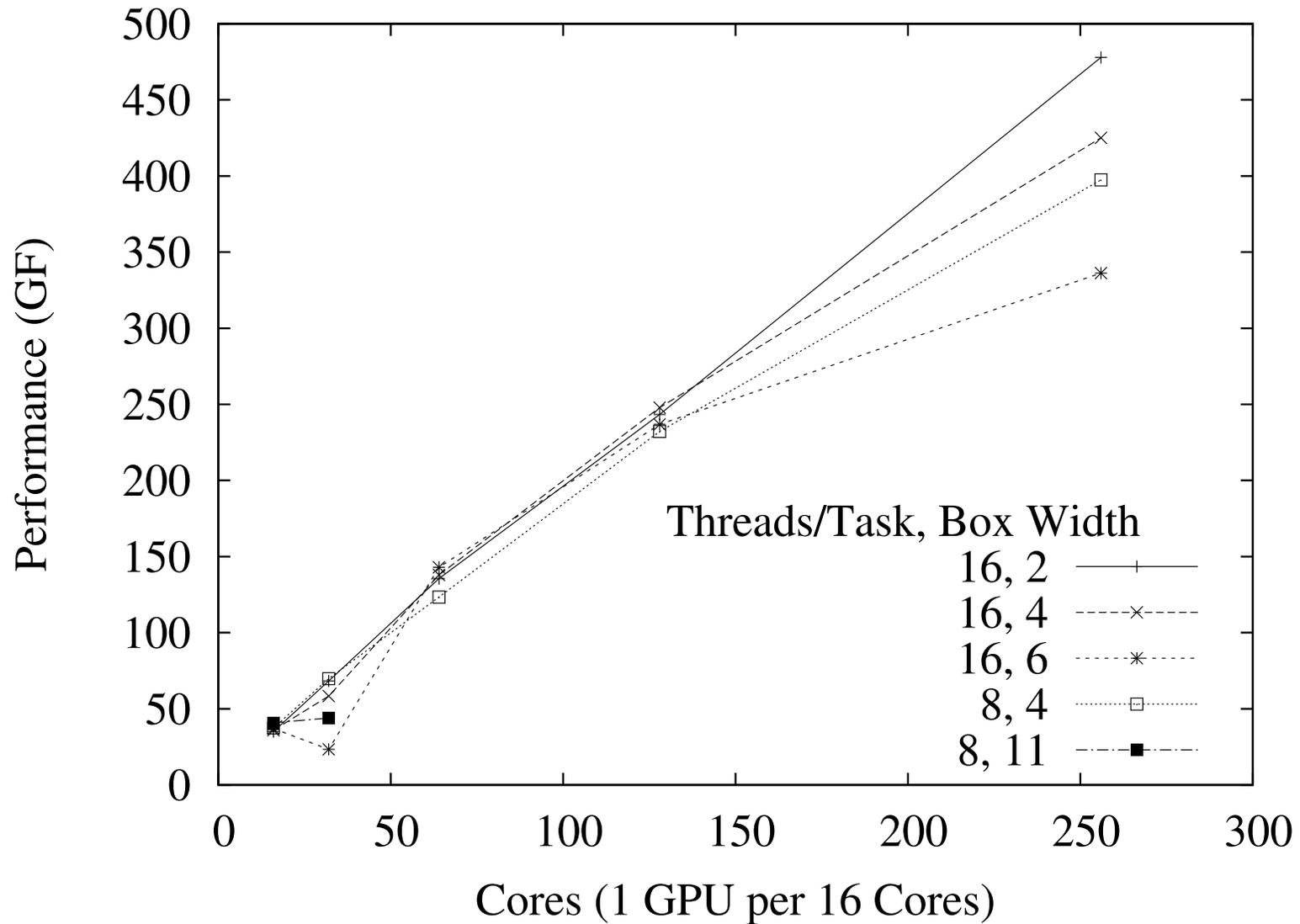
# Best Performance on Yona



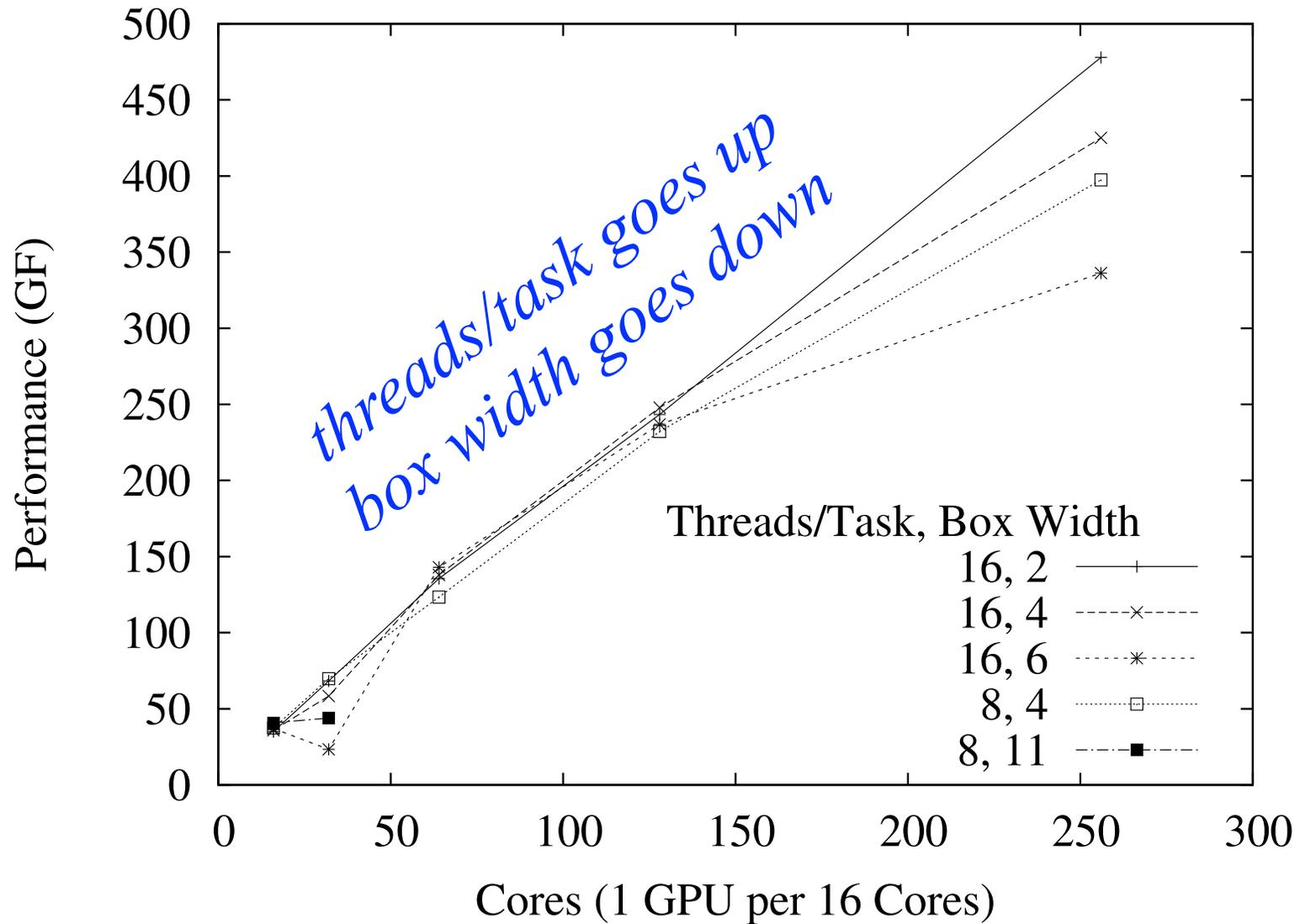
# Best Performance on Yona



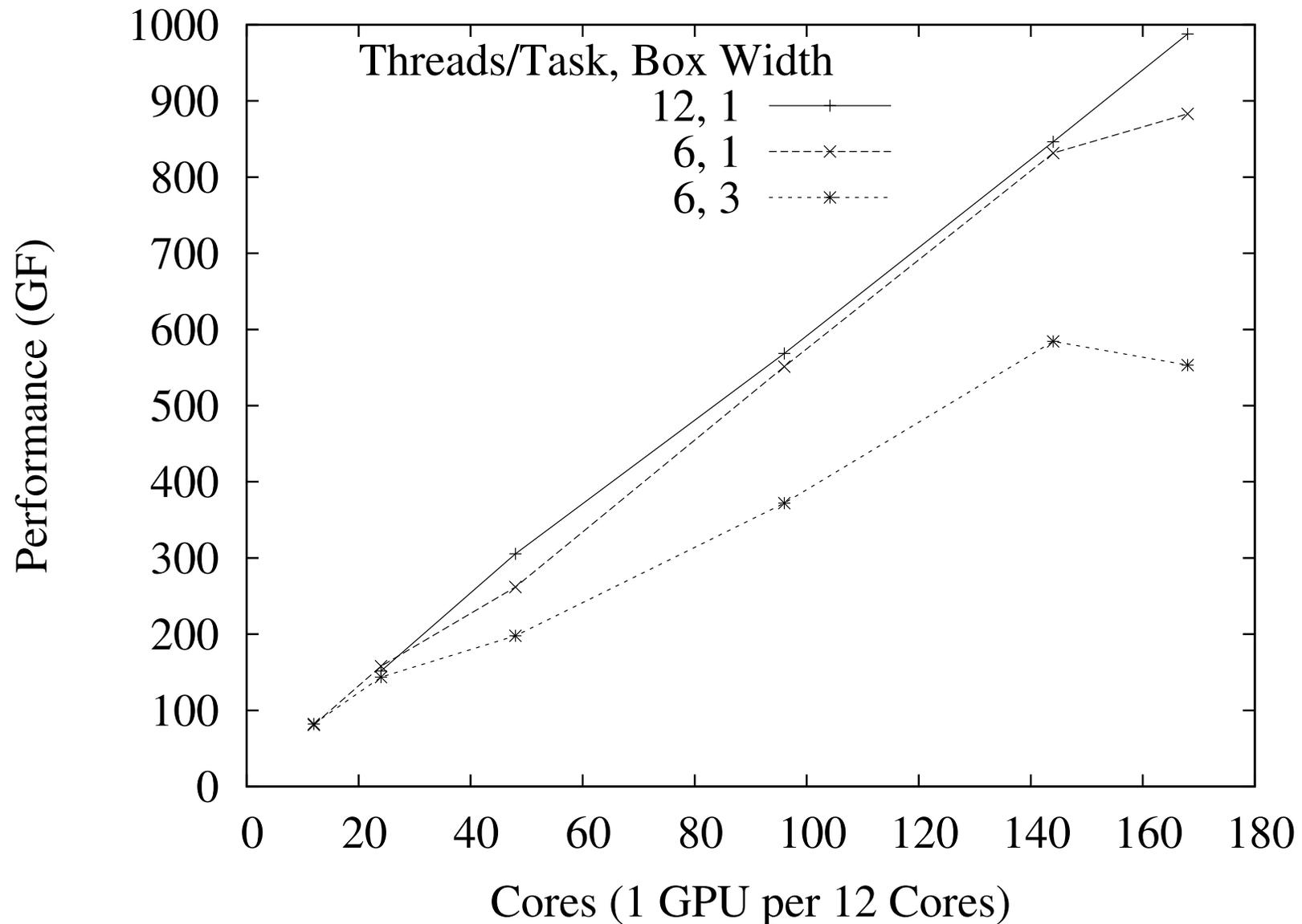
# CPU-GPU Overlap Performance on Lens



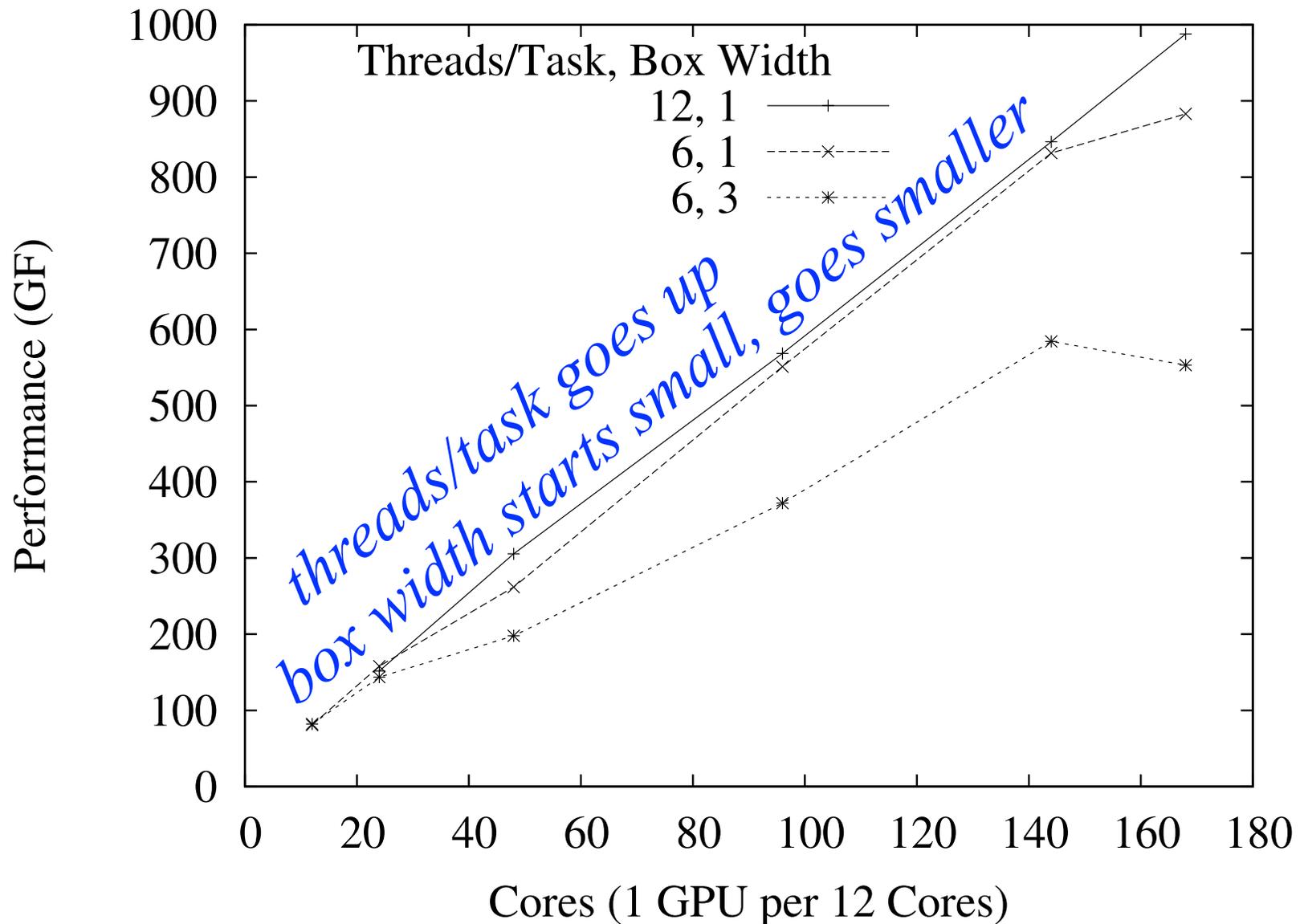
# CPU-GPU Overlap Performance on Lens



# CPU-GPU Overlap Performance on Yona



# CPU-GPU Overlap Performance on Yona



# Overlapping Computation and Communication for Advection on Hybrid Parallel Computers

- MPI overlap less important for this test
- But tuning threads/task *is* important
- Overlapping CPU computation, GPU computation, MPI communication, and CPU-GPU communication
  - Improves performance by more than 2x
  - Matches GPU-resident performance per GPU
- Best performance from giving minimal (but *non-vanishing*) work to CPU
- Performance comes at a 4x cost in lines of code

# Overlapping Computation and Communication for Advection on Hybrid Parallel Computers

James B White III (Trey)

trey@ucar.edu

National Center for Atmospheric Research

Jack Dongarra

dongarra@eecs.utk.edu

University of Tennessee, Knoxville

Programming Weather, Climate, and Earth-System Models  
on Heterogeneous Multi-Core Platforms

NCAR, September 8, 2011

based on work first presented at IPDPS, Anchorage, AK, May 17, 2011

*Portions of this work were funded by the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research, both of the US Department of Energy. This research used resources of the OLCF at Oak Ridge National Laboratory and of NERSC at Lawrence Berkeley National Laboratory, both of which are supported by the Office of Science of the US Department of Energy.*