PIO-PNETCDF-Restart benchmarks

Benchmarks for the PIO-PNETCDF Restart code used by the CAM physics package and the HOMME dycore. These numbers should be similar to what PIO/PNETCDF can achieve for CAM history when that code is finished.

Methodology

Setup CAM/HOMME for aqua planet simulations (see Running CAM-HOMME)
Set restart_option = 'end' in the drv_in namelist.

Code changes

Add instrumentation to PIO calls by adding -DTIMING to USER_CPPDEFS line in Makefile

Output

- ne120nv4 case (1/4 degree average grid spacing at the equator), 26 levels.
- restart file: 12,370 MB
- Runs on 128 processors (~1.5GB per processor)
- BG/P 512 nodes vn mode (512mb per core): wrote PIO restart files, but ran out of memory on surface restart files.
- BG/P 512 nodes smp mode (2gb per processor) ?
- BG/P 1024 nodes vn mode ?

Results

- homme_cam3_6_19 branch
- NCPUS: number of cores (MPI threads)
- io_cpus: PIO num_iotasks
- stripe: number of Lustre OST's the file is stripped across
- All times in seconds
- MB/s computed from pio_write_nf() time. (does not include re-arranger or other CAM and PIO overhead)

SNL Blackrose (intel/openmpi/infiniband linux cluster, Lustre filesystem)

NETCDF			
NCPUS/io_cpus /stripe	**cam_write_restart**	**pio_write_nf**	**MB /s**
128/128/64	170.4	151.5	82
128/128/16	149.5	128.2	96
128/128/4	183.9	168.5	75
128/128/1	333.5	317.7	40
128/32/32	149.8	143.9	
128/32/8	144.2	138.5	

In the NETCDF case, the difference between cam_write_restart and pio_write_nf is mostly due to the data re-arranger.

Parallel NETCDF			
NCPUS/io_cpus /stripe	**cam_write_restart**	**pio_write_nf**	**MB /s**
128/128/128	663.4	385.4	32
128/128/64	485.8	121.0	102
128/128/32	844.5	601.8	21
128/32/32	146.7	98.5	126
128/8/16	222.5	174.0	71
128/8/8	156.3	99.9	124

In the PNETCDF case, sometimes the calls to "pio_put_var_0d_int" take a significant amount of time, ~200s.

ORNL Jaguar Cray XT4 with Lustre

ANL BG/P with GPFS

Note: for comparison, the standalone HOMME dycore on BG/P can write restart files using MPI-I/O directly. On 8192 cores, writing a 22.8GB restart file:

- MPI collective with a derived type: 7.2s (3.2 GB/s)

- Asynchronous, non-overlapping MPI_File_write_at(): 8.5 MB/s (ouch!).

NETCDF			
NCPUS /io_cpus	**cam_write_restart**	**pio_write_nf**	**MB /s**
2048/2048	207.8	201.6	61
2048/128	137.7	136.1	91

Parallel NETCDF			
NCPUS/io_cpus	**cam_write_restart**	**pio_write_nf**	**MB /s**
8192/2048	86.0	16.9	732
2048/2048	71.1	19.4	638
2048/512	37.8	20.8	595
2048/128	41.9	32.7	