# Parallelism in non-spectral, global GKW

- For global version, radial wavevectors replaced by finite differences and multi-point gyro-average:
- Significant changes to field solve & gyroaverage

$$G = G^\dagger = J_0(k_\perp(\mu)\rho_s)$$

local spectral

$$(G(\mathbf{x},\mu,\mathrm{sp})f(\mathbf{x}))_i = G_{ij}f_j$$

non-spectral

$$\sum_{\mathrm{sp}} \int \mathrm{d}^3\mathbf{v} \left[ Z_s e G f(\mathbf{x}) + F_M \frac{Z_s^2 e^2}{T_s}(GG^\dagger - 1)\phi(\mathbf{x}) \right] = 0$$

$$I(k_y, x_i, s) + P_{ij}\phi(k_y, x_j, s) = 0$$

- Advantages: Global profiles, another direction for domain decomposition:
  - 6D distribution $f(k_y, x, s, \mu, v_\|, \mathrm{sp})$
  - 3D fields $\phi(k_y, x, s)$
  - 5D gyro-averaged fields $<\phi>(k_y, x, s, \mu, \mathrm{sp})$
  - 4D collisions conservation field
  - 5D domain decomposition $x, s, \mu, v_\|, \mathrm{sp}$
  - 5D MPI Cartesian communicator $x, s, \mu, v_\|, \mathrm{sp}$

# Field solve becomes complex, multi-stage MPI

<u>Non-spectral parallel field solve</u>

Local integral → 5D field $J_{\ell v_\parallel} = \int_{\ell v_\parallel} \mathrm{d}v_\parallel Z f$

↓

MPI P2P (5D field, 1D com): x buffers $J_{\ell x} \to J_{Lx}$

↓

Gyro-av., local int. → 3D field $I_{\ell_{\mu,v_\parallel,\mathrm{sp}}} = \int_{\ell\mu} \mathrm{d}\mu\, G J_{\ell v_\parallel} \frac{2\pi B}{m}$

↓

MPI Allred. sum (3D field, 3D com) $I_{g_{\mu,v_\parallel,\mathrm{sp}}} = \Sigma_{g\mathrm{sp},gv_\parallel,g\mu} I_{\ell_{\mu,v_\parallel,\mathrm{sp}}}$

↓

MPI Allgather x (3D field, 1D com) $I_{\ell x,gy} \to I_{gx,\ell y}$ y scatter

(would need double P2P even if polarisation solve not global)

↓

**Implicit** solve (UMF), **global** in x $\phi_{gx,\ell y} = -P^{-1}_{gx,\ell y} I_{gx,\ell y}$

↓

MPI Allgather y (3D field, 1D com) $\phi_{gx,\ell y} \to \phi_{\ell x,gy}$ x scatter

↓

MPI P2P (3D field, 1D com): x buffers $\phi_{\ell x} \to \phi_{Lx}$

↓

Gyro-av. (store in 5D) $< \phi >_i (\mathbf{x}, \mu, \mathrm{sp}) = G_{ij}(\mu, \mathrm{sp})\, \phi_j(\mathbf{x})$

- Communicate radial buffer cells before each gyroav.
- Polarization solve implicit
  - Matrix ~ $Nx^2$
  - Uses UMFpack library
  - No radial decomposition
  - Instead parallelise over binormal modes, using Cartesian direction usually used for velocity
  - Allows good scaling
  - MPI Allgather / AllScatter
- 5 MPI communications total
  - Only Allreduce is costly
- Limitation: Gyroradii may cross <=1 proc. boundary

# Field solve becomes complex, multi-stage MPI

| Non-spectral parallel field solve | | **28.9** |
|---|---|---|
| Local integral → 5D field | $J_{\ell v_\parallel} = \int_{\ell v_\parallel} \mathrm{d}v_\parallel Z f$ | 2.9 |
| MPI P2P (5D field, 1D com): x buffers | $J_{\ell x} \longrightarrow J_{Lx}$ | 1.3 |
| Gyro-av., local int. → 3D field | $I_{\ell_{\mu,v_\parallel,\mathrm{sp}}} = \int_{\ell\mu} \mathrm{d}\mu\, G J_{\ell v_\parallel} \frac{2\pi B}{m}$ | 3.8 |
| MPI Allred. sum (3D field, 3D com) | $I_{g_{\mu,v_\parallel,\mathrm{sp}}} = \Sigma_{g\mathrm{sp},gv_\parallel,g\mu} I_{\ell_{\mu,v_\parallel,\mathrm{sp}}}$ | 5.4 |
| MPI Allgather x (3D field, 1D com)<br>(would need double P2P even if polarisation solve not global) | $I_{\ell x,gy} \longrightarrow I_{gx,\ell y}$ y scatter | 4.5 |
| **Implicit** solve (UMF), **global** in x | $\phi_{gx,\ell y} = -P^{-1}_{gx,\ell y} I_{gx,\ell y}$ | 7.3 |
| MPI Allgather y (3D field, 1D com) | $\phi_{gx,\ell y} \longrightarrow \phi_{\ell x,gy}$ x scatter | 1.3 |
| MPI P2P (3D field, 1D com): x buffers | $\phi_{\ell x} \longrightarrow \phi_{Lx}$ | 0.3 |
| Gyro-av. (store in 5D) | $<\phi>_i (\mathbf{x}, \mu, \mathrm{sp}) = G_{ij}(\mu, \mathrm{sp})\, \phi_j(\mathbf{x})$ | 0.2 |

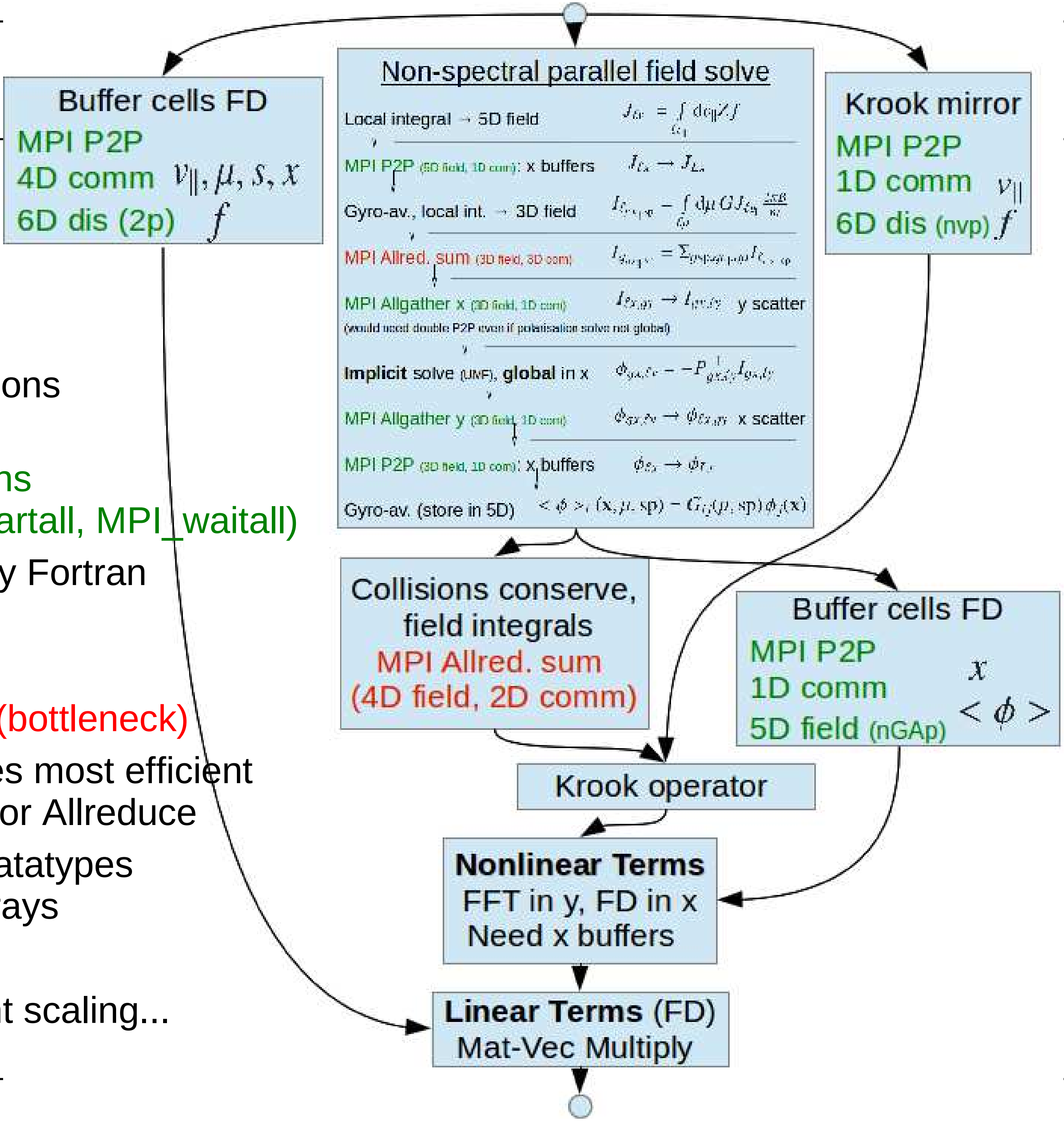
Percentages of total runtime for large electromagnetic collisional case on 32,768 processes

Important for good MPI scaling that the runtime fraction of the field solve does not rise

# All MPI ops

<u>Single Explicit Step</u>

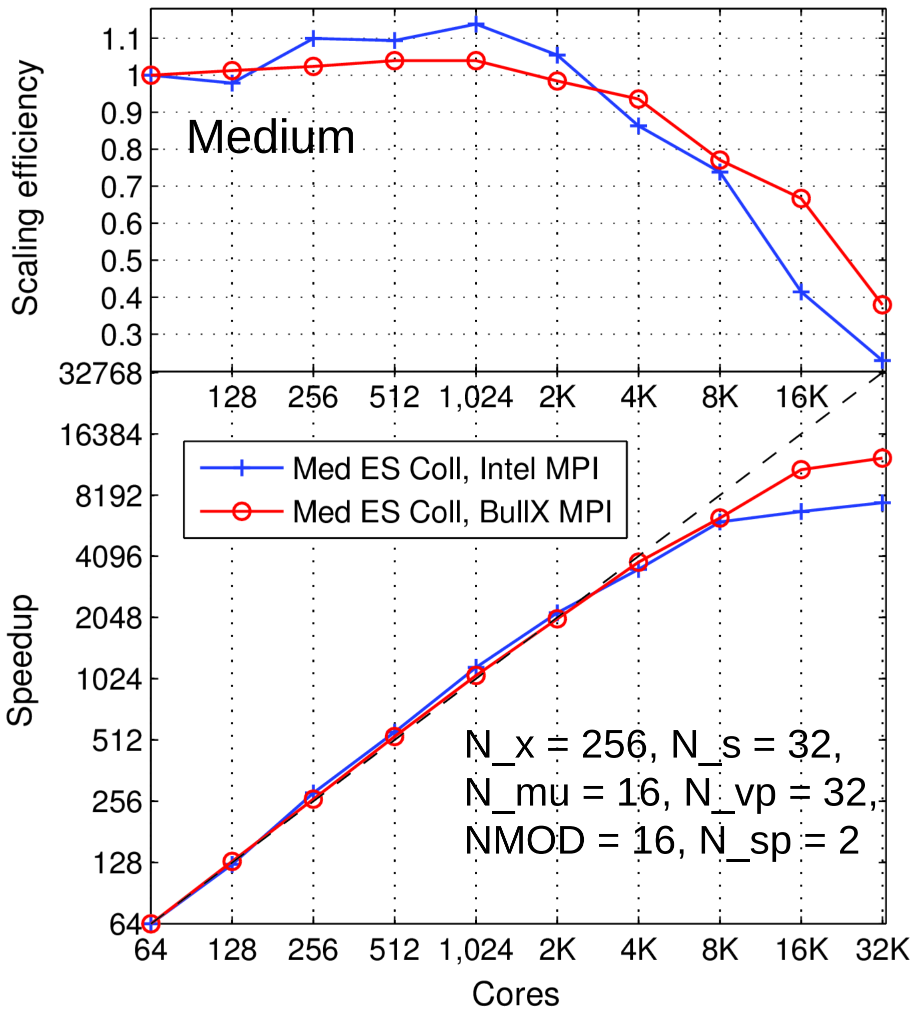
- Overlap communications with computation
- 7 P2P communications nonblocking (MPI_Startall, MPI_waitall)
  - Each managed by Fortran structure
- 2 MPI Allgather ops
- 2 MPI Allreduce ops (bottleneck)
- Parallel direction gives most efficient scaling: No Integrals or Allreduce
- 12 x 2 MPI derived datatypes for different buffer arrays
- All this allows efficient scaling...



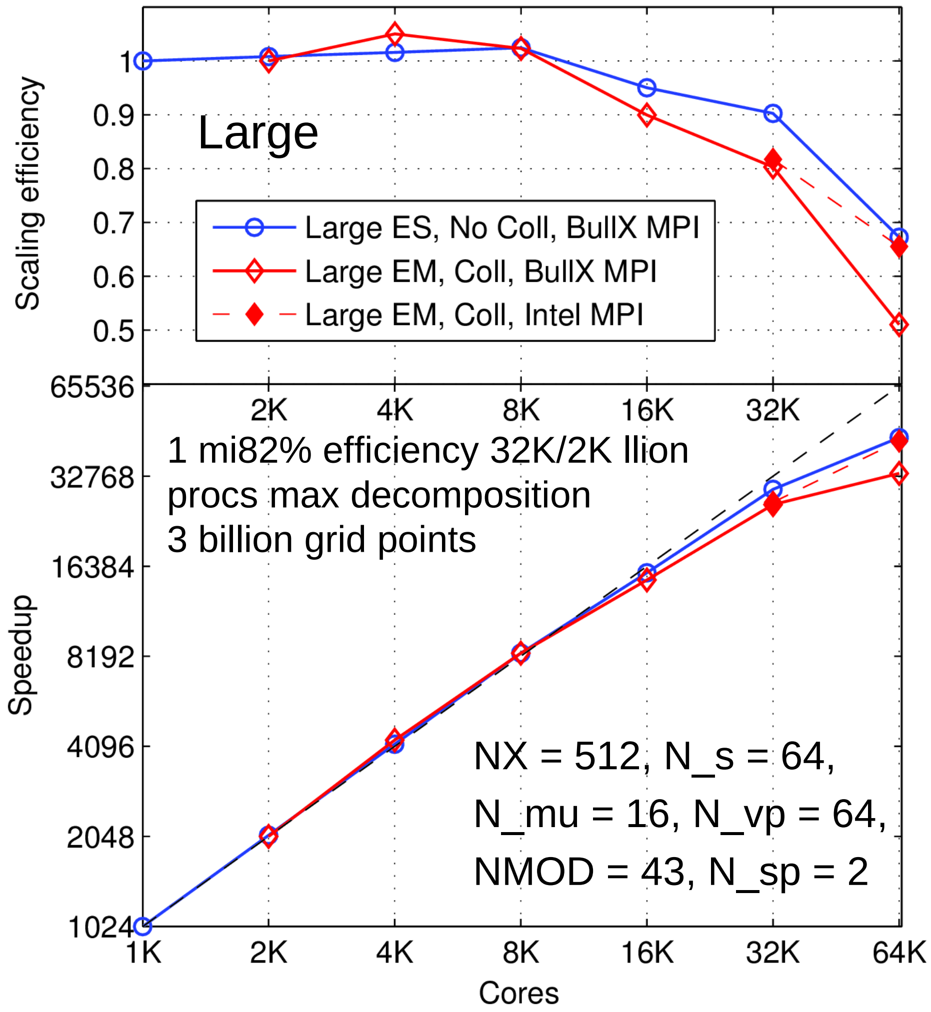
# Strong Scaling Performance (HELIOS)

For typical physics problems: electromagnetic collisions (no conservation, no Krook)

95% efficiency 4K / 64




82% efficiency 32K / 2K



# Comments on the HELIOS scaling

- Omit timing of the first loop (order of seconds for first MPI calls)

- Scaling breaks down when field solve fraction of runtime increases
  - Inefficient Collectives: large Allreduce and Allgather
  - Use MPI_STATUSES_IGNORE, MPI_IN_PLACE

- Intel MPI and BullX MPI perform differently
  - Intel MPI manages cache better affects memory bandwidth ?
  - Intel MPI better at largest collectives
  - BullX good at smaller # of cores
  - Many environment variables to tweak for largest jobs
  - No general optimal settings
    - Each case / code is different
    - Intel MPItune might help ?
    - OMP_COLL_OPT

| Large, time (s): | Bull MPI | | Intel MPI | |
|---|---|---|---|---|
| 1600 iterations | 32K | 64K | 32K | 64K |
| Field Solve | 21 | 25 | 26 | 22 |
| Non Linear | 13 | 07 | 12 | 06 |
| Buffer cells f | 09 | 11 | 10 | 05 |
| Linear Term | 21 | 11 | 17 | 08 |
| **Total** | **70** | **56** | **68** | **43** |
| **Efficiency** | **0.81** | **0.52** | **0.82** | **0.66** |

# Summary (MPI part)

- Explicit Eulerian gyro-kinetics with finite differences on appropriate grids allows parallel domain decomposition in 5 dimensions

- Permits efficient exploitation of the largest present day supercomputers
  - Limited by available resources, not code capability

- Requires complex multi-dimensional MPI
  - multiple communicators
  - multiple derived datatypes
  - persistent non blocking communications
    (some person-years of development and testing)

- Allows us to do new and interesting physics
  (e.g. NTM simulations in first part of the talk)