Present and Future Computing Requirements for Tokamak Edge Physics

C.-S. Chang

Princeton Plasma Physics Laboratory

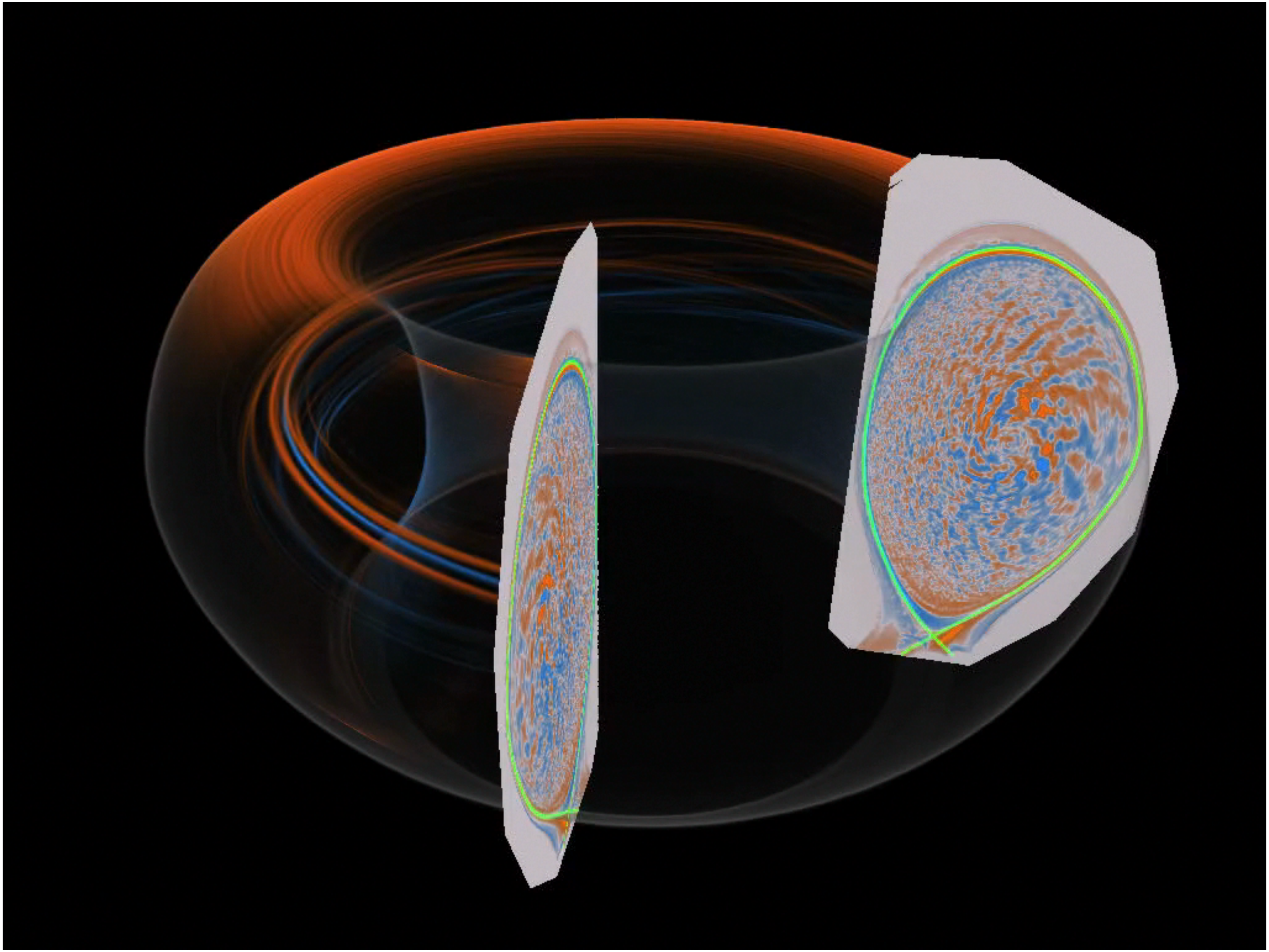
NERSC BER Requirements for 2017
September 11-12, 2012
Rockville, MD

1. Project Description

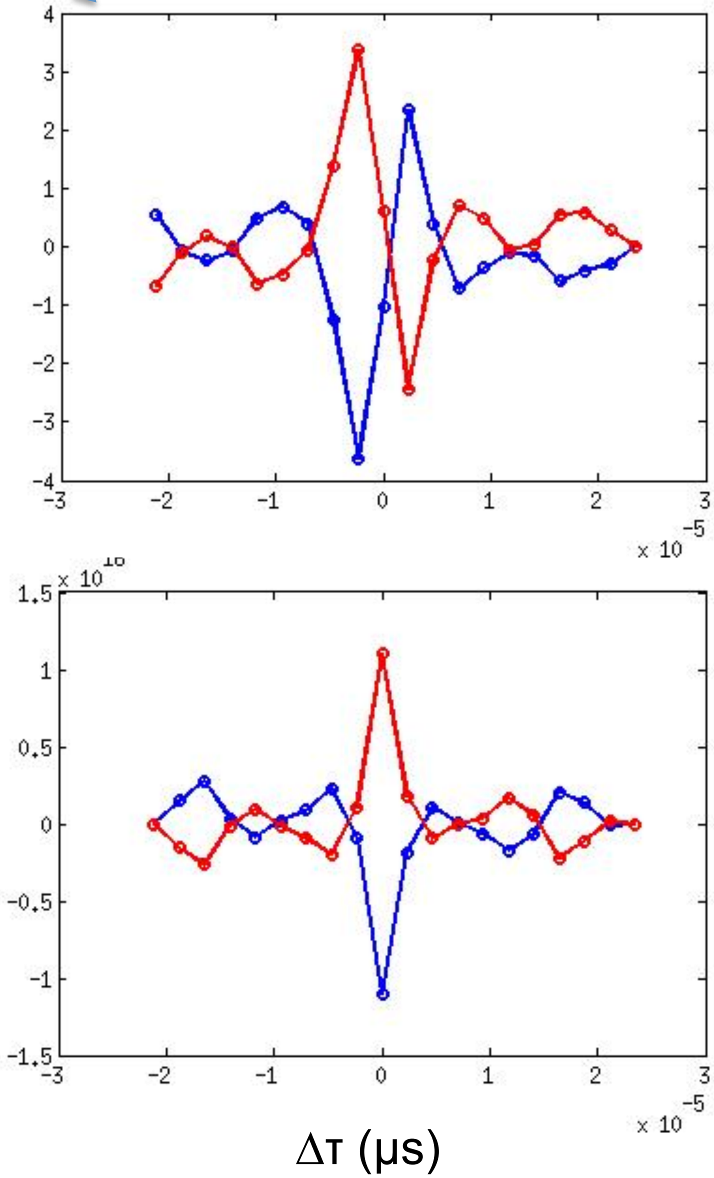
C.-S. Chang, PPPL and SciDAC EPSI (XGC family codes)
X. Xu, LLNL (TEMPEST/COGENT and BOUT++)

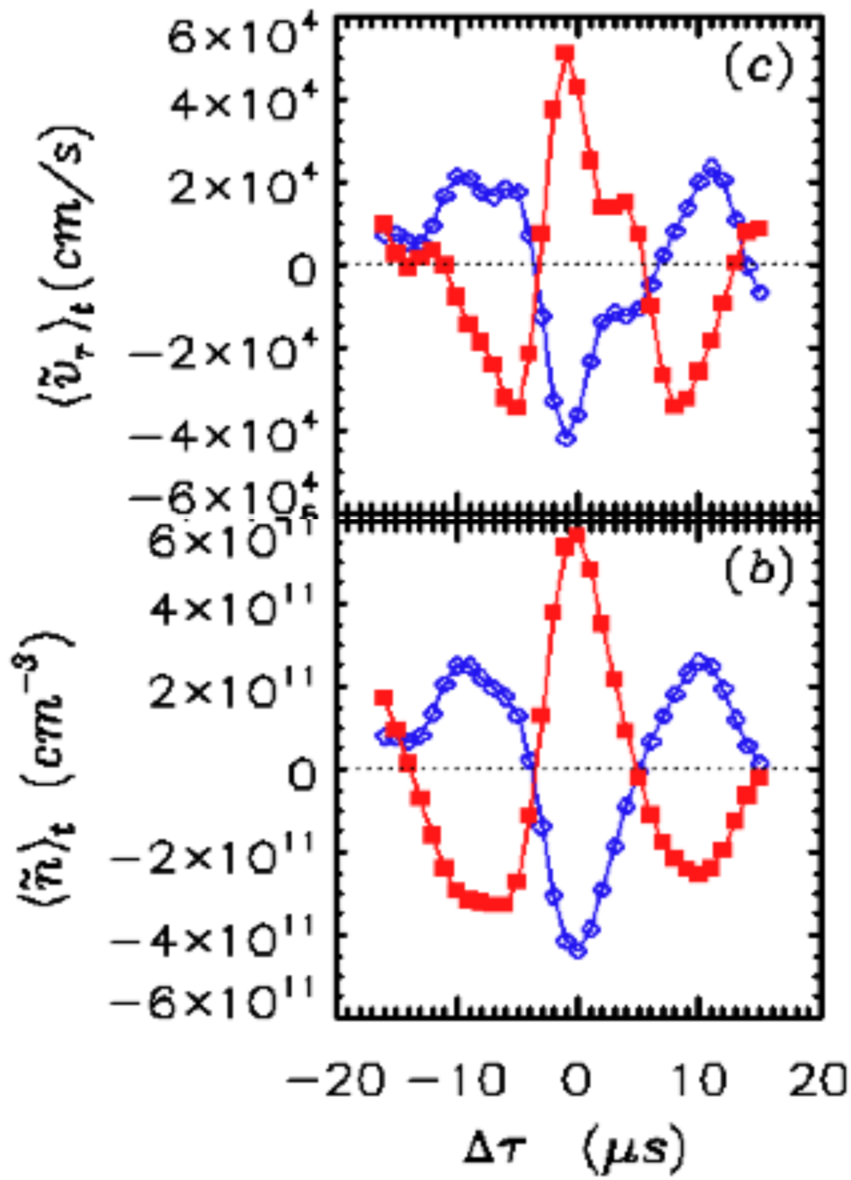
XGC1: Kinetic PIC (ODE) code in diverted geometry

- Our present focus is in
  - Adding continuum technology into PIC method → Total-f
  - Electrostatic turbulence, background evolution, ions, electrons, neutral particles, impurity particles: Built-in DEGAS-2
  - 3D perturbed magnetic field from XGC0 (→XGCa)
  - Adding in electromagnetic micro-turbulence
- By 2017 we expect to have
  - Full-scale gyrokinetic electromagnetic turbulence with both parity
  - Higher fidelity PMI data (in collaboration with SciDAC PWI)
  - Higher fidelity atomic physics (ADAS)
  - 3D physics built into XGC1, consistently with E&M turbulence
  - Experimental edge time scale simulation in ITER

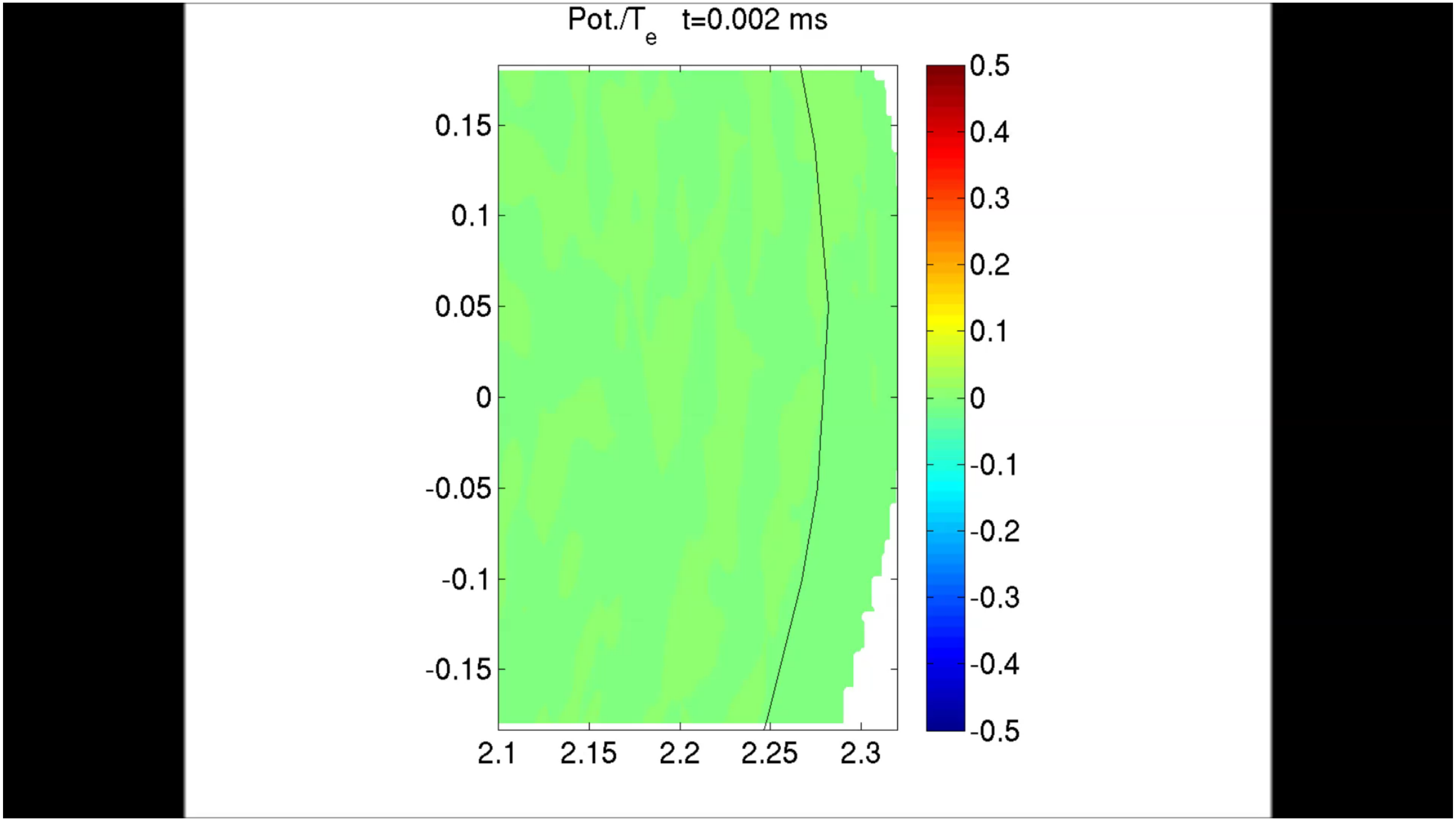


XGC1 comparison of non-local ITG turbulence with experiment (Tynan's group) is encouraging [DIII-D, in the pedestal region]

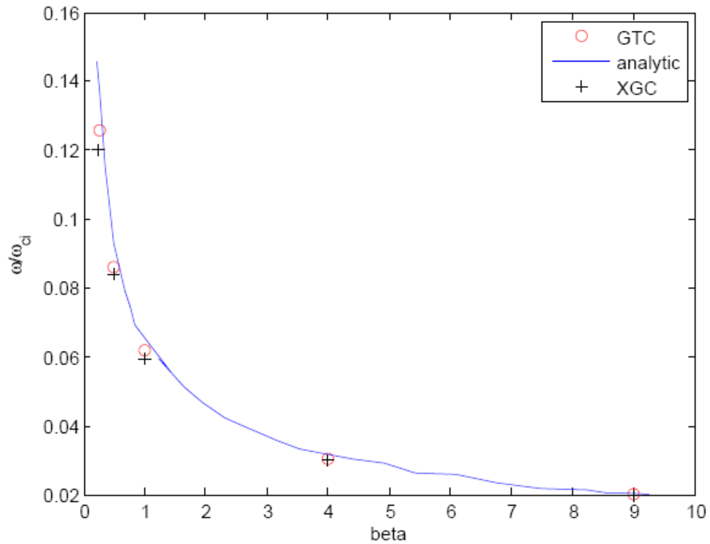




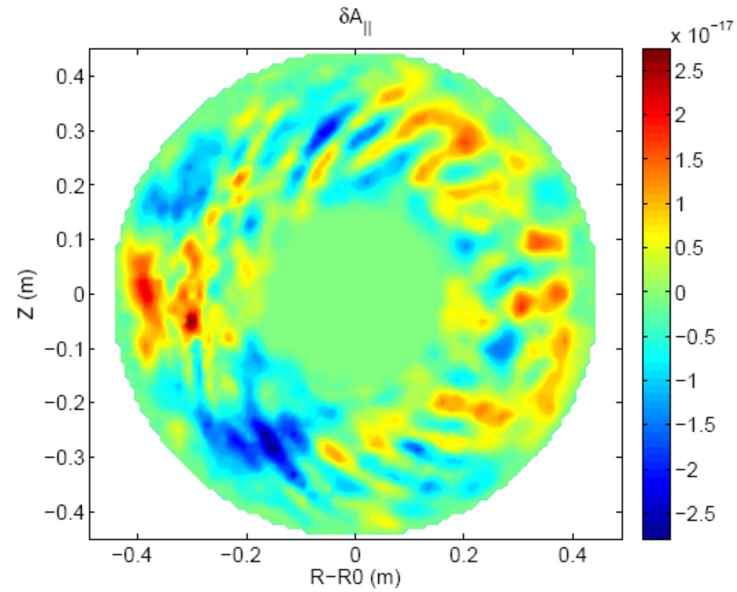
Kinetic electrons produce non-linear coherent structures ("blobs") across the separatrix surface



Electromagnetic turbulence capability in XGC1



Cross verification of fluid Alfven wave in fluid-particle hybrid scheme



Low beta ITG turbulence in electromagnetic split-weight scheme

- Currently two E&M turbulence methods are being developed
  - fluid-particle hybrid method
  - δf split-weight scheme
- These two methods are being converted into total-f in diverted geometry
- PPPL's double split scheme and German improved cancellation scheme (Bottino, Hatzky et al) will be examined.

XGC0

- Neoclassical PIC transport code in diverted geometry
  - 3D Lagrangian equation of motion: drift kinetic code
    - Kinetic ions, kinetic electrons, kinetic impurity particles
  - Monte-Carlo neutrals with wall recycling, and impurity radiation
  - Toroidally and poloidally averaged electrostatic potential solver
  - Mimic turbulent transport as random and directional particle walk
  - 3D perturbed B-field penetration and pedestal response

XGCa

  - Gyrokinetic neoclassical code: performs gyro-averaging
  - Solver includes poloidal potential variation
- Present focus
  - Higher fidelity neoclassical physics in scrape-off layer & separatrix
- By 2017 we expect to have
  - Multiscale coupled simulation with XGC1 for experimental edge-time scale simulation of ITER (~50ms in DIII-D)

TEMPEST/COGENT

- Full-function kinetic continuum (PDE) code
- Present focus
  - Solve 4D axisymmetric physics including rf heating from CQL3D
- By 2017 the aim is to have rf and neutral beam heating cability with added-on anomalous transport

BOUT++: Fluid code in diverted geometry

- LLNL-York collaborative code. Users at KSTAR and EAST.
- Present focus
  - Addition of gyro-Landau effect with nonlocal transport coefficients
- By 2017 we expect to have
  - Integrated simulation with TEMPEST/COGENT
  - Magnetic islands and plasma-wall interaction
  - Simulation of multiple ELM cycle

2. Computational Strategies

XGC family codes

- We approach this problem computationally by
  - Full-function particle simulation on largest HPCs
  - Realistic diverted tokamak geometry with material wall boundary
  - All the important multiscale kinetic physics in one executable
- The codes we use are
  - XGC1 gyrokinetic turbulence-neoclassical code
  - XGC0 drift-kinetic neoclassical flux-surface averaged solver code
  - XGCa gyrokinetic neoclassical axisymmetric solver code
- XGC1 is characterized by these algorithms:
  - Small scale dynamics is on particle space using Lagrangian ODE
  - Large scale dynamics is put on the coarse-grained 5D X-V space
  - Fully nonlinear integro-differential Fokker-Planck-Landau collsion operator is solved on 5D X-V space using finite difference method, and mapped back to the particle space
    - Linear based Monte-Carlo collision operator is an option

XGC family codes (2)

- Our biggest computational challenges are
  - Reducing inter-node communications and intra-node communications between GPU and CPUs → smart domain-decomposition
  - Reducing error in the long wave length E&M tearing modes
- Our parallel scaling is limited by
  - Inter-node communications in the kinetic electron simulation → smart domain-decomposition
  - We do not see scalability problem with ions
- We expect our computational approach and/or codes to change (or not) by 2017 in this way
  - XGC1 has already been ported to heterogeneous CPU-GPU
    - 30X GPU compared to one CPU core
    - 2X with 16cores+1GPU compared to 32 cores: apple-to-apple → Improve
  - Smart-domain decomposition for faster electron communication
  - Multiscale coupling between XGC1 and XGCa for experimental edge time-scale simulation of ITER

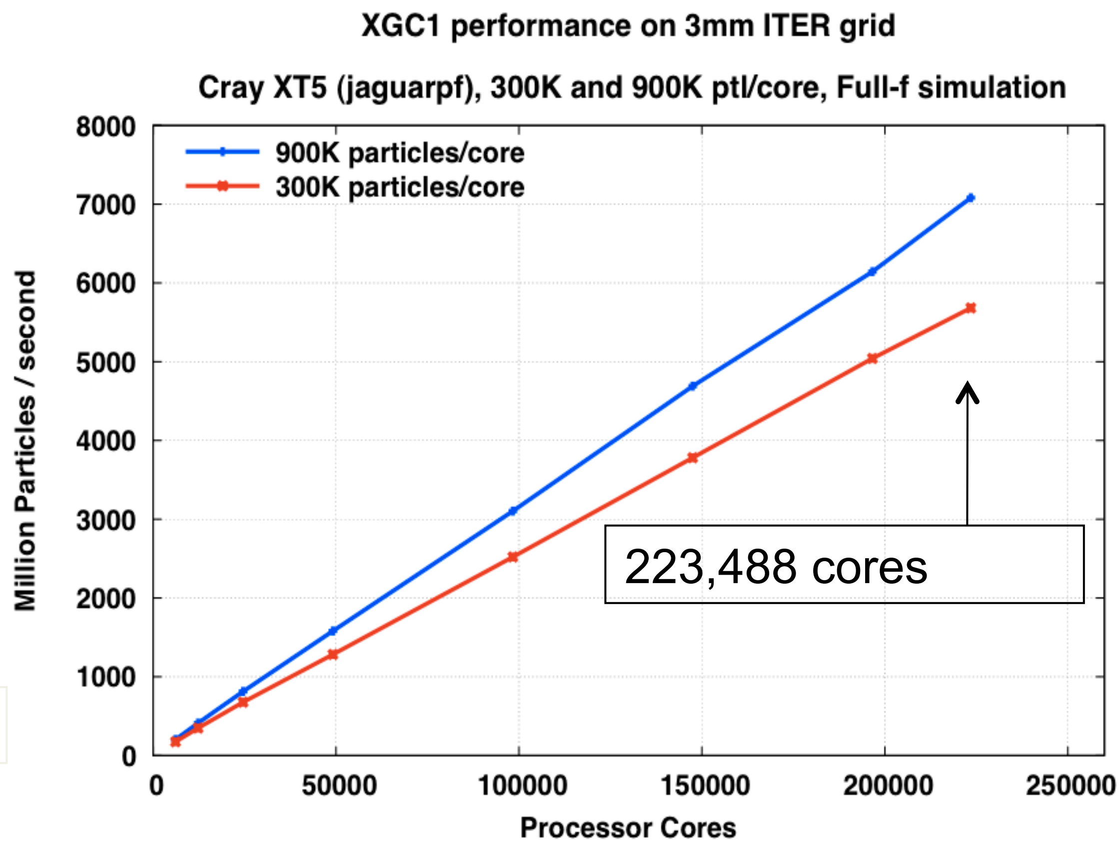
XGC1 scalability on Jaguar

XGC1 is a modern extreme-scale gyrokinetic code

- Ion scaling is linear to the maximal number of Jaguar nodes; multi domains
- One day result is targeted. More physics or larger system → more number of particles → proportionally more number of cores.

Smaller tokamak devices fall on approximately the same lines due to smallness of the solver time (~5%).

Hybrid MPI-OpenMP



BOUT++

- Our biggest computational challenges are
  - Fraction of time spent in communication rapidly increases beyond 2K cores
- Our parallel scaling is limited by
  - "Size of each MPI subdomain becomes smaller with increased concurrency, resulting in more boxes with a fewer number of grid points per MPI box. This results in an increase of the surface to volume ratio (at the highest concurrency of 65,536 there are only two grid points in the poloidal direction)."
- We expect our computational approach and/or codes to change (or not) by 2017 in this way
  - Similar approach will continue

3. Current HPC Usage

XGC family codes

- Machines currently using
  - Hopper and Carver at NERSC, Titan at OLCF
  - PPPL clusters for new routine development
- Hours used in 2012
  - 60M CPU hours on Hopper
- Typical parallel concurrency and run time, number of runs per year
  - 73,728 cores on Hopper, with XGC1 and with the real-mass electron version XGC0: only limited by Queue wait time
    - 20 hours run time, about 30 total runs per year including restarts
  - 1,000 cores on Hopper with ion version XGC0
    - 20 hours run time, 500 runs per year (parameter scan)
- Data read/written per run
  - 6TB, mostly checkpoint files from XGC1

XGC family codes (2)

- Memory used per (node | core | globally)
  - 5GB/node, 0.208GB/core, 20TB globally
- Necessary software, services or infrastructure
  - Adios, PETSc, MPI, OpenMP
  - MATLAB for data analysis on Carver
  - Consulting and account services
- Data resources used (HPSS, NERSC Global File System, etc.) and amount of data stored
  - 300TB on HPSS, 22% used up as of 3/17/2013
  - 1TB on NERSC Global File System
  - 20TB on Scratch Space

TEMPEST/COGENT

- Currently not running on Hopper
  - – Uses a local LLNL cluster

BOUT++

- Typical parallel concurrency and run time, number of runs per year
  - – Typically use 512 to 1024 cores for a single job for 4 to 8 hours
  - – Multiple restarts to get the physics results
- Hours used in 2012: 2.6M hours around the globe
- Necessary software, services or infrastructure
  - Combo, LAPACK, Hypre, Sundials, mppl, Python, pyMPI, Numpy, swig

4. HPC Requirements for 2017

XGC family codes

- Compute hours needed (in units of Hopper hours)
  - Can use 2B Hopper hours at 32X Hopper capability, for ITER
- Changes to parallel concurrency, run time, number of runs per year
  - Node level parallel concurrency will remain the same, assuming that the number of nodes stay at ~10,000 level
  - Intra-node parallelism will improve withGPUs and CPUs (XGC1 is already on GPU-CPU)
  - The 1-day run-time requirement remains the same
  - Number of runs remain similar to execute 32X greater problem size per run
- Changes to data read/written
  - More in-situ on-on-the-fly data analysis to reduce the written data
  - On-the-fly in-memory data compression/decompression for faster check-pointing

XGC family codes (2)

- Changes to memory needed per (core | node | globally)
  - Per node: 50GBminimum, 100GB preferred (XGC codes low memory codes) → not much more than present node memory
  - Globally: 500TB minimum, 1PB preferred
- Data resources needed (HPSS, NERSC Global File System, etc.) and amount of data to be stored
  - 3000TB on HPSS,10X of 2013 requirement
  - 10TB on NERSC Global File System
  - 600TB on Scratch Space: 30X
- Changes to necessary software, services or infrastructure
  - Faster inter- and intra-node communication speed is key requirement at 32X power

TEMPEST/COGENT

- Changes to parallel concurrency
  - Requires 2,048 processors for a single ITER simulation

BOUT++

- Changes to parallel concurrency
  - 4,096 processors for 96 hours

Together (TEMPEST/COGENT and BOUT++)

- Total compute hours needed in 2017: 100M hours
- Average number of runs: 350 runs/year
- Which is more important to your project: strong scaling or weak scaling?: Strong scaling
- Changes to necessary software, services or infrastructure

  "Once a large joint code development is finished, the team is disassembled, physicists typically still remain to maintain and run the codes. Should NERSC provides the necessary supports?"

5. Strategies for New Architectures

- **XGC**: Strategy for running on new many-core architectures (GPUs or MIC)
  - We have already ported XGC1 to heterogeneous GPU-CPU system (Todi and Titan). Our codes are highly adaptable to GPU-CPU.
  - MIC can also be used as well
- **XGC**: To date we have prepared for many core by
  - Using hybrid MPI-OpenMP
  - XGC1 with gyrokinetic ions runs to the maximal number of Jaguar cores (223,488) with linear scalability. 2D particle domain decompostion and 1D grid domain decomposition are used.
  - Enhanced communications by fast kinetic electrons, without any electron domain decomposition, spoil this linear extreme scalability. → Smart domain decomposition

- **XGC**: We are already planning to do
  - Further optimization of the heterogeneous GPU-CPU computing
  - A smart electron domain decomposition to improve the extreme scale scalability to maximal Titan capability (~20 PF)
  - In-situ, on-the-fly data management using Adios DataSpaces technology
    - Including on-the-fly, in-memory data compression
  - Further improvement of the asynchronous adaptable I/O technology
  - Collaboration with the SciDAC SUPER, SDAV, FASTMath Institute is established (and QUEST)
- **XGC**: To be successful on many-core systems we will need help with
  - Close collaboration with NERSC consultants on networking, compiling, debugging, other software, and systems

5. Summary

- What new science results might be afforded by improvements in NERSC computing hardware, software and services?
  - **XGC**: GPUs could enable larger scale kinetic electron physics
  - **TEMPEST/COGENT and BOUT++**: Integrated modeling
- Recommendations on NERSC architecture, system configuration and the associated service requirements needed for your science
  - **XGC**: Fast inter and intra node network
  - **XGC:** We have our own **eSiMon** Dashboard service based upon Adios. However, using the NERSC **Gateway** service for analysis and collaboration on **extreme scale data** can be beneficial on the large jobs run on extreme scale NERSC machines. Synergistic use between eSiMon and Gateway services can be studied.

- NERSC generally refreshes systems to provide on average a 2X performance increase every year. What significant scientific progress could you achieve over the next 5 years with access to 32X your current NERSC allocation?
  - **XGC**: It can enable us to perform all-scale electromagnetic edge physics simulation in realistic diverted ITER geometry with neutrals, impurities, wall recycling and sputtering, over experimental edge time scale (~50ms)
  - **TEMPEST/COGENT and BOUT++**: Multiscale multiphysics
- What "expanded HPC resources" are important for your project?
  - **XGC**: GPUs (or MIC)
  - **XGC**: Larger scale data analysis capability, connected to Gateway services