

Improving the Performance of Convolutions

Many algorithms exhibit poor data locality when translated directly from the underlying mathematical formulation, so that naive programming, without detailed understanding of the underlying processor architecture, yields unsatisfactory results. But careful reorganization can produce computations performing at nearly the peak processor theoretical maximum.

We show an example of this with of 2D and 3D anisotropic convolutions. These important algorithms, which have no exploitable computational symmetries, run at less than 20% of peak using either the gcc or pathscale compilers. A rethinking of the data usage to enhance locality, combined with careful tiling of the data arrays, produces 3D performance at near 90% of peak on an sgi Octane. These performances have been shown to scale over thousands of cores on a SiCortex 5832.

2D and 3D Anisotropic Convolutions

- Samara Technology Group, LLC (STG) developed a 2D and 3D Anisotropic Convolution library for SiCortex MIPS-based supercomputer.
- Requirements: scale evenly to thousands of processors, while demonstrating near peak performance on single processors.
- These routines, now part of the STG scientific library, have been ported with similar results to Nehalem, Opteron, PPC970 and ARM.
- Rough grained approach is for optimal 2D routine to process “slices” in 3D space, retaining memory locality. Slices are summed over the appropriate number of vertical 3D taps to produce output points.
- Fine grained approach is to load smaller 2D kernel “box” into registers, make initial pass on a 2D tile and perform partial sums for subsequent “boxes.”
- Outputs are calculated in tiles so that memory is only read/written one time. This property, particularly important for scaling, has been verified using the STG Performance Technology Platform (PTP).

Anisotropic Convolutions are the most general, and hence the most complicated to calculate, of all the convolutions. They are frequently employed in seismic, physics and other scientific applications requiring very large scale HPC systems.

Convolution Comparisons on MIPS

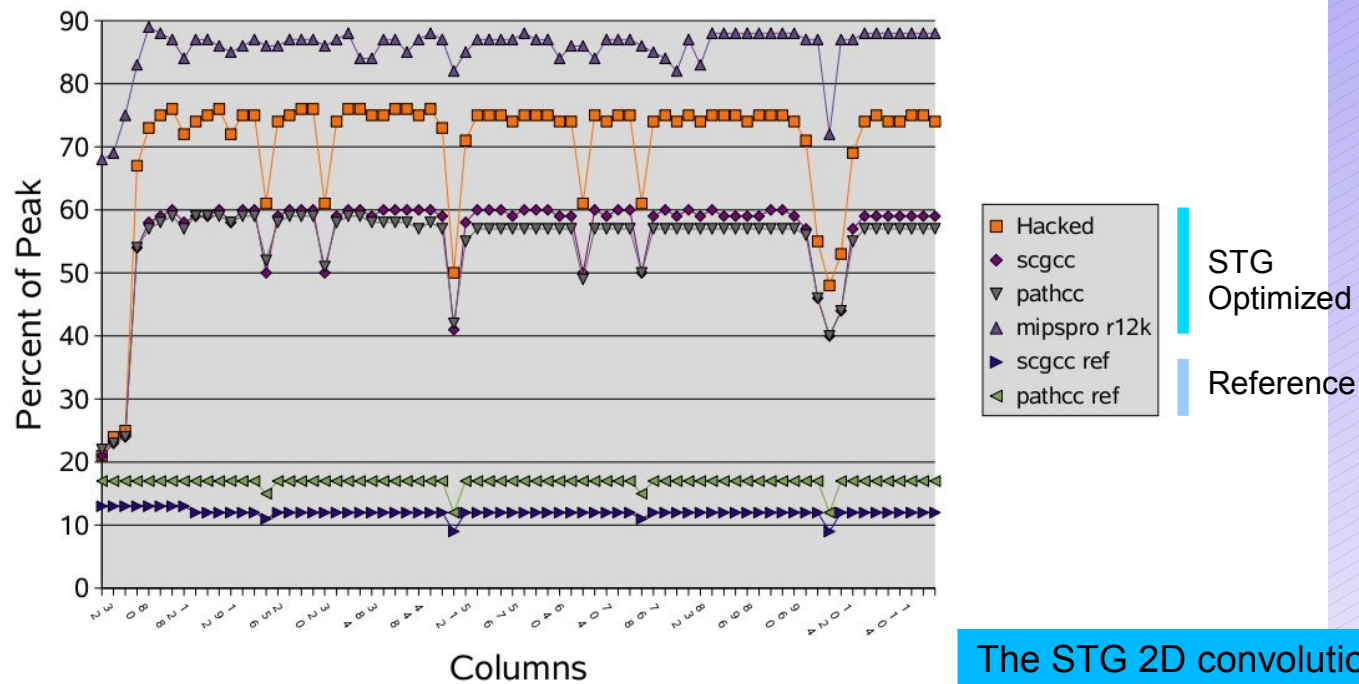
- Compare the STG highly local algorithm with the reference algorithm in several ways:
 - “hacked”: the STG algorithm compiled with gcc and then manually modified to implement software pipelining.
 - “scgcc”: the unmodified gcc compilation of the STG algorithm using “best found options.”
 - “pathcc”: the STG algorithm compiled with the pathscale compiler using “best found options.” (This compiler did not then implement software pipelining.)
 - “scgcc ref”: the reference algorithm compiled with gcc under high optimization.
 - “pathcc ref”: the reference algorithm compiled with the pathscale compiler under high optimization.
 - “mipspro r12k”: the STG algorithm compiled using the MIPSPRO compiler (software pipelining implemented) for an sgi r12K Octane target.

2D Convolutions

- 15 x 15 tap kernel:
 - SiCortex rev A system peak achieves 76% of theoretical maximum system performance
 - Octane r12K system peak achieves 89% of theoretical maximum system performance
 - Best reference performance achieves 17% of theoretical maximum system performance
- 33 x 33 tap kernel:
 - SiCortex rev A system peak achieves 81% of theoretical maximum system performance
 - Octane r12K system peak achieves 93% of theoretical maximum system performance
 - Best reference performance achieves 21% of theoretical maximum system performance

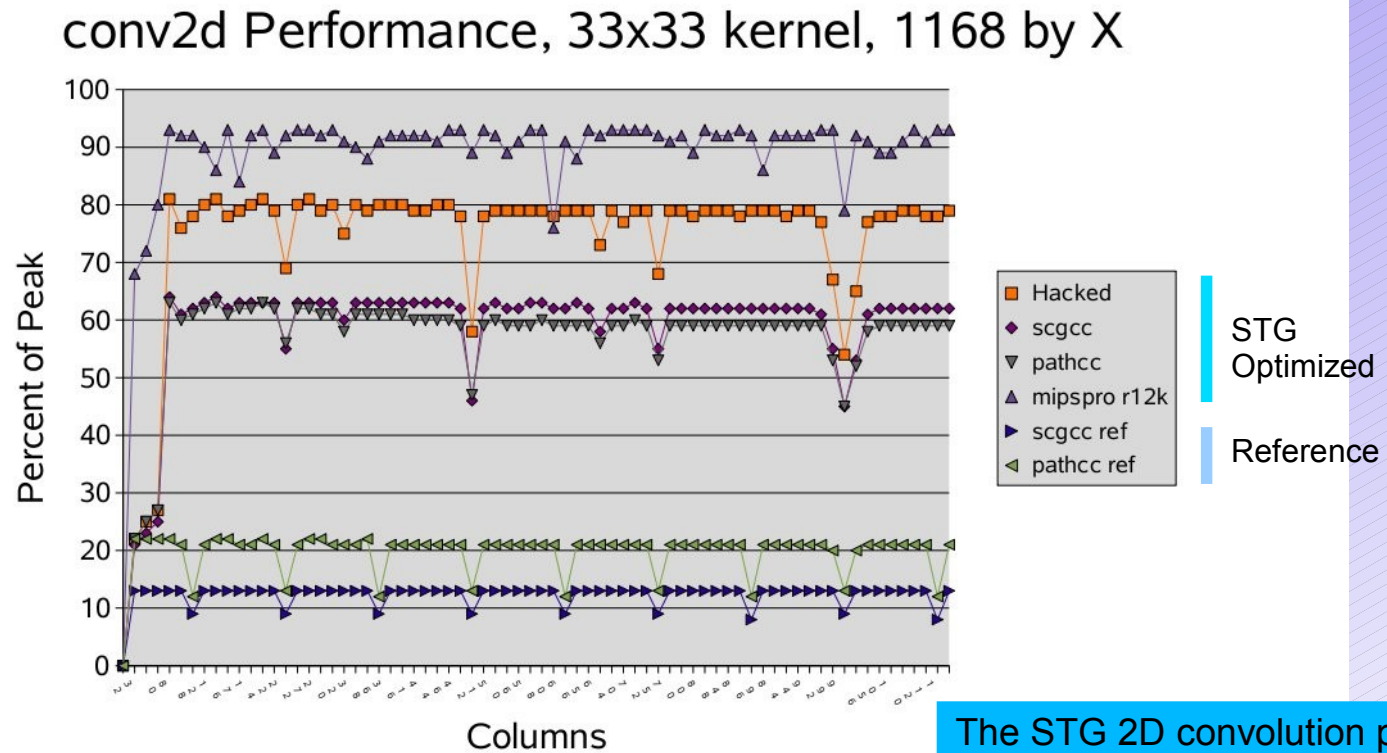
15 x 15 2D Anisotropic Convolution, varying column size

conv2d performance, 15x15 kernel, 1168 by X



The STG 2D convolution performs at 4.5 X the reference algorithm; and it scales across thousands of nodes.

33 x 33 2D Anisotropic Convolution, varying column size

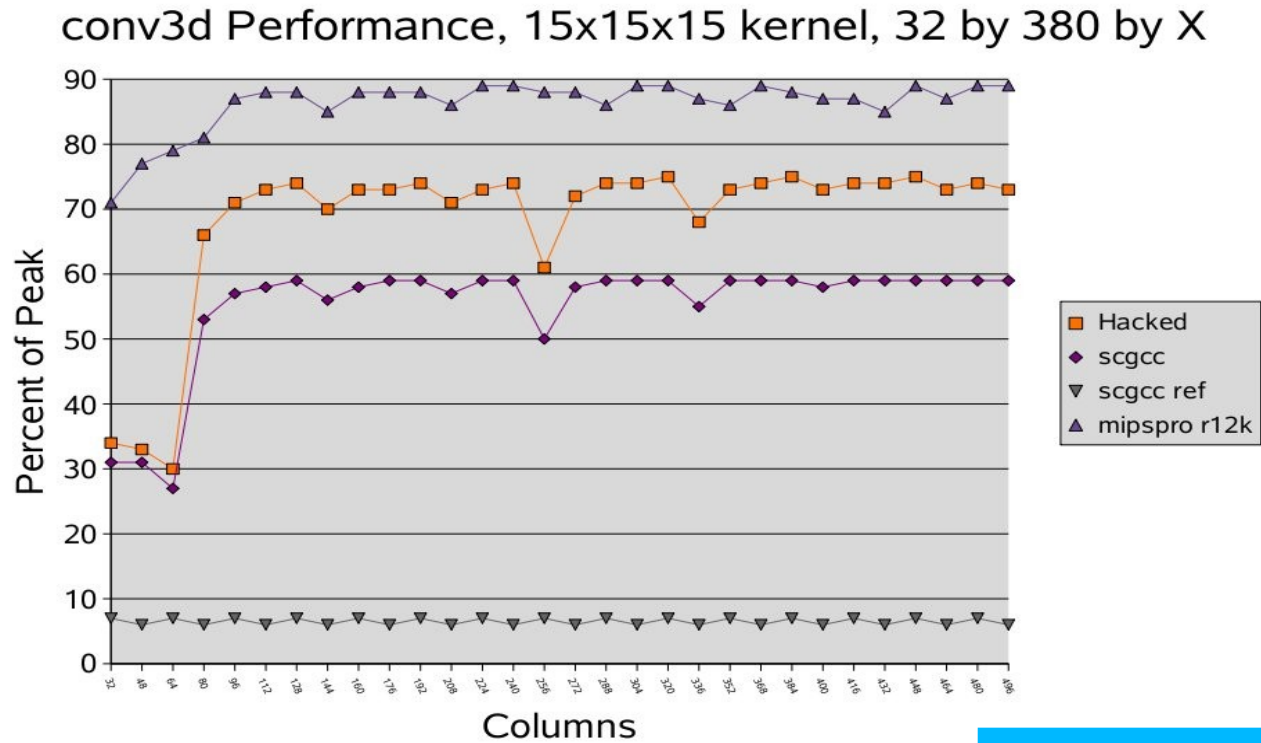


The STG 2D convolution performs at 4 X the reference algorithm with near perfect scaling even across thousands of nodes.

3D Convolutions scale from 2D

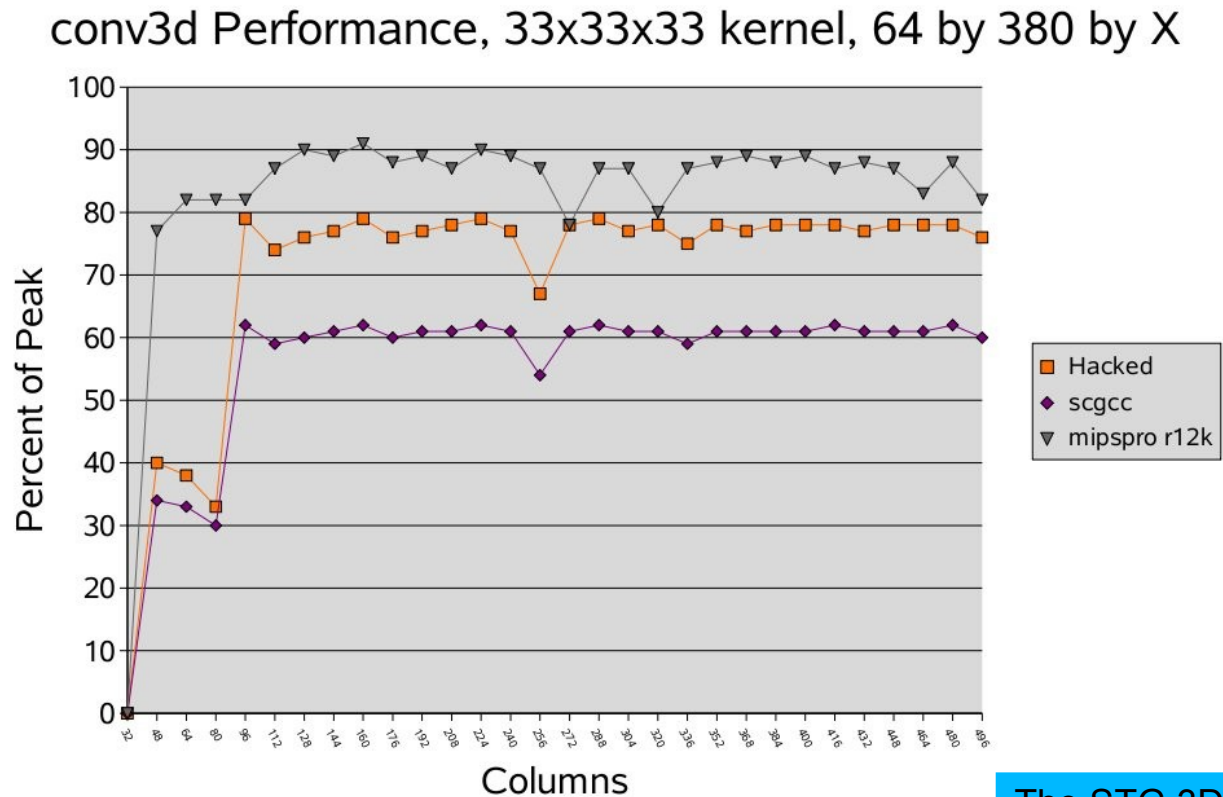
- 15 x 15 x 15 tap kernel:
 - SiCortex rev A system peak achieves 75% of theoretical maximum system performance
 - Octane r12K system peak achieves 89% of theoretical maximum system performance
 - Best reference performance achieves 8% of theoretical maximum system performance (only half of 2D performance)
- 33 x 33 x 33 tap kernel:
 - SiCortex rev A system peak achieves 79% of theoretical maximum system performance
 - Octane r12K system peak achieves 91% of theoretical maximum system performance

15 x 15 x 15 3D Anisotropic Convolution, varying column size



The STG 3D convolution performs at 9.5X the reference algorithm on the SiCortex Ice9A (MIPS 5kf).

33 x 33 x 33 3D Anisotropic Convolution, varying column size



The STG 3D convolution sustains nearly 90% of r12K theoretical maximum across different sizes.