OpenTuner: An Extensible Framework for Program Autotuning

Hauptseminar WS 2016/2017

Matthias Tietz

10.02.2017

Gliederung

1. Einleitung

- Suche nach optimaler Programm-Implementierung teil-/vollautomatisierbar
- bestehende Autotuning-Frameworks meist domänenspezifisch (ATLAS, FFTW)

OpenTuner:

- Ausdrücken des Suchraums statt direkter Optimierung
- Verfahren komfortabler und oft effizienter (große Suchräume)
- mehrere Optimierungsziele komibinierbar
- stark auf Erweiterbarkeit ausgelegt

verwendete Begriffe:

- **Parameter**: Variablen eines bestimmten Typs
- **Konfiguration**: konkrete Zuweisungen von Parametern
- **Konfigurations-Manipulator**: Bearbeiten der Parameter in Konfig.

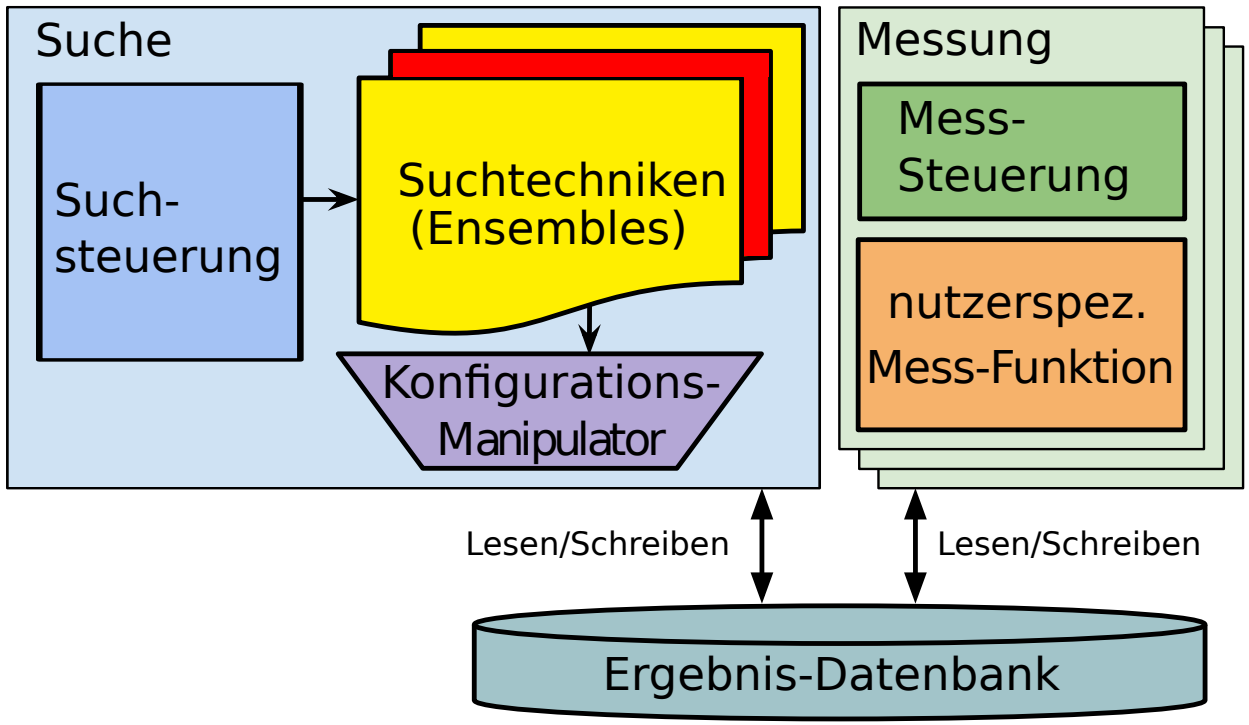
- **Suchraum**: zu durchsuchende Menge von Konfigurationen
- **Suchtechnik**: Suchraum erforschen (KM), Anfragen für Messung (Ergebnis)
- **Suchprozess**: Lesen/Schreiben von Konfigurationen

- **Messprozess**: konkrete Konfiguration ausführen
- **Messresultat**: Ergebnis der Messung, abhängig von Optimierungsziel(en)
- **Ergebnisdatenbank**: Festhalten aller Ergebnisse während Tuning-Vorgangs

Herausforderungen bei der Entwicklung eins Autotuning-Frameworks:

- 1. Eine passende Konfigurations-Repräsentation
  - Darstellung der domänenspezif. Datenstrukturen und Bedingungen
  - Qualität der Repräsentation entscheidend für Effizienz des Autotuners
- 2. Größe des gültigen Konfigurations-Raumes
  - Kürzen des Konfigurations-Raumes $\rightarrow$ Verlieren guter Lösungen
  - rießige Konfigurationsräume möglich $\rightarrow$ intelligente Suchtechniken notwendig
- 3. Beschaffenheit des Konfigurations-Raumes
  - Suchräume in der Praxis meist sehr komplex
  - domänenspezif. Suchtechniken notwendig

2. OpenTuner Framework



Verwendung

- 1. Suchraum definieren (Konfig.-Manipulator)
- 2. `run()`-Methode definieren: Auswerten der Konfig. im Suchraum $\rightarrow$ Ergebnis
- 3. Festlegen des Optimierungsziels
- Umsetzung mittels kleinem Python-Programm (OpenTuner API)

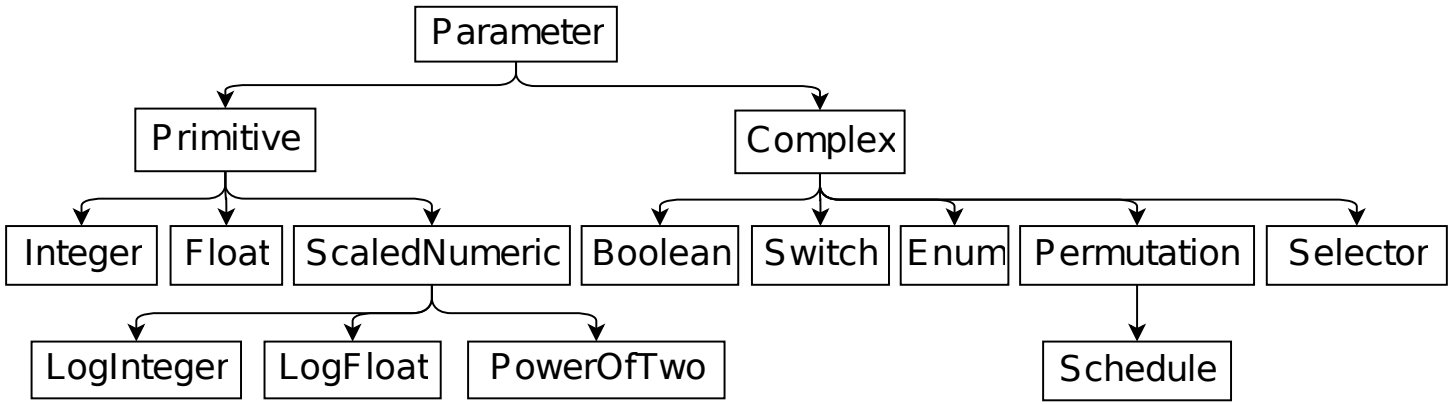
Suchtechniken

- OpenTuner stellt Suchtechniken für viele Suchraum-Typen bereit
- Ausführen mehrerer Suchtechniken gleichzeitig (Ensembles)
- dynamische Testzuweisung anhand Erfolges dieser Techniken

Konfigurations-Manipulator

- Abstraktionsschicht zwischen Suchtechnik und Konfigurations-Struktur
- Zugriff auf Parameter-Objekte: Lesen/Schreiben per Suchtechnik

Parameter-Typen



Primitive Parameter

- numerische Werte mit Unter-/Obergrenze
- `Float` und `LogFloat (-Int)` gleiche Repräsentation in Konfiguration, aber zugrundeliegender Wert für Suchtechnik skaliert
- Grund: ohne Logskal. würde Effekt der Wertänderung mit steigender Parametergröße sinken

Komplexe Parameter

- haben variables Set an Manipulatoren $\rightarrow$ stochastische Parameter-Änderungen
- `Boolean, Switch` und `Enum` bewusst als komplex. Parameter, Darstellung als ungeordnete Sammlung $\rightarrow$ es existiert kein Gradient (wie bei prim. Param.)
- `Permutation` : Liste von Werte inkl. Manipulatoren zur randomisierten Änderung der Reihenfolge
- `Schedule`, `Selector`

Optimierungsziele

- OpenTuner unterstützt mehrere Ziele, default: `time`
- `accuracy, energy, size, confidence` oder ein nutzerdef. Ziel
- Es können auch mehrere Ziele zugleich verfolgt werden, bspw. *Genauigkeit einhalten, gleichzeitig Zeit minimieren*

3. Konkrete Anwendungen der Entwickler

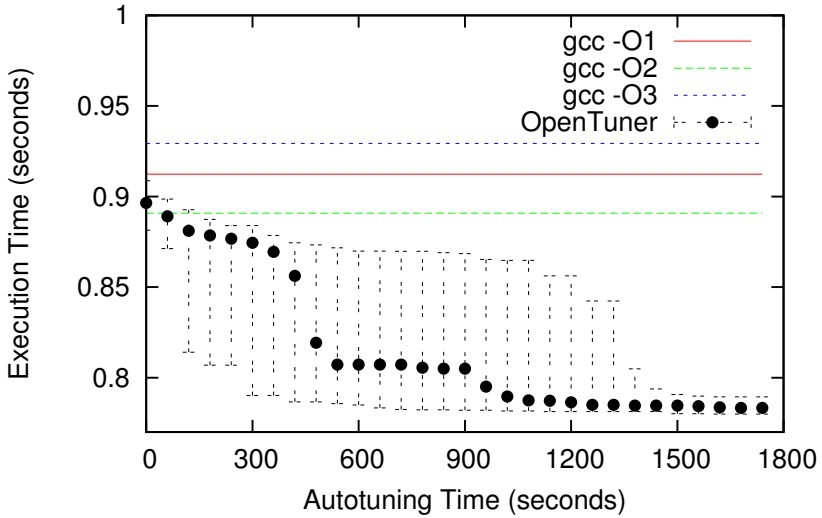
GCC/G++ Flags

- klassische Parameter-Optimierung
- unterstützte Flags: `g++ --help=optimizers`
- Parameter inkl. zulässiger Wertebereiche: `params.def` (gcc source code)
- Implementierung des Autotuners:
  - 1. Erstellen des `configuration manipulator`
  - 2. Erstellen der `run`-Funktion
  - 3. Festlegen des Optimierungsziels

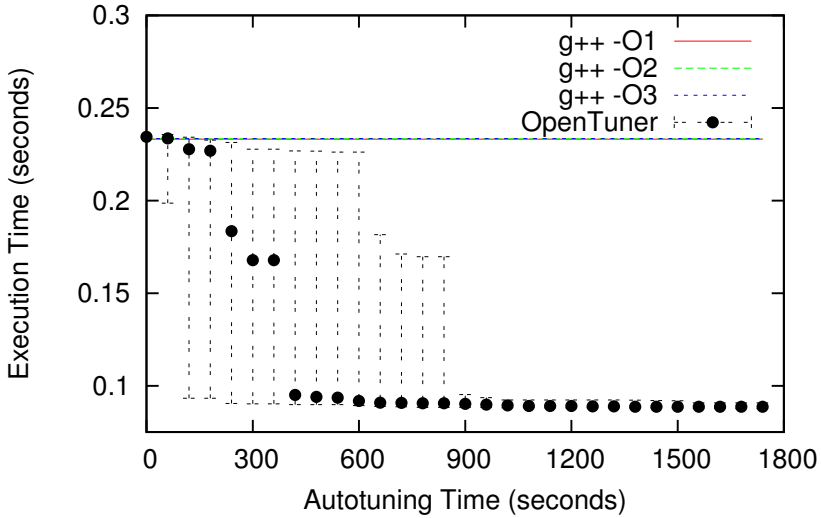
```
import opentuner
from opentuner import ConfigurationManipulator
from opentuner import EnumParameter
from opentuner import IntegerParameter
from opentuner import MeasurementInterface
from opentuner import Result

GCC_FLAGS = [
  'align-functions', 'align-jumps', 'align-labels',
  'branch-count-reg', 'branch-probabilities',
  # ... (176 total)
]

# (name, min, max)
GCC_PARAMS = [
  ('early-inlining-insns', 0, 1000),
  ('gcse-cost-distance-ratio', 0, 100),
  # ... (145 total)
]
```

```
class GccFlagsTuner(MeasurementInterface):

  def manipulator(self):
    """
    Define the search space by creating a
    ConfigurationManipulator
    """
    manipulator = ConfigurationManipulator()
    manipulator.add_parameter(
      IntegerParameter('opt_level', 0, 3))
    for flag in GCC_FLAGS:
      manipulator.add_parameter(
        EnumParameter(flag,
                      ['on', 'off', 'default']))
    for param, min, max in GCC_PARAMS:
      manipulator.add_parameter(
        IntegerParameter(param, min, max))
    return manipulator
```

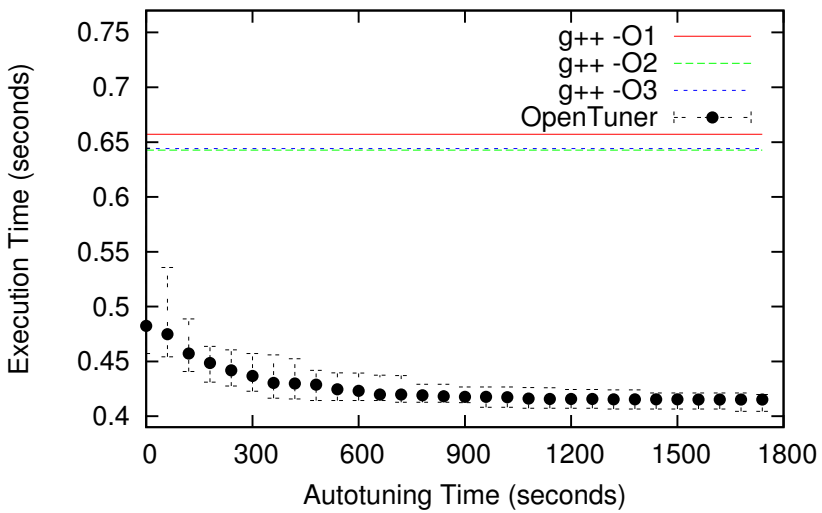
```
def run(self, desired_result, input, limit):
  """
  Compile and run a given configuration then
  return performance
  """
  cfg = desired_result.configuration.data
  gcc_cmd = 'g++ raytracer.cpp -o ./tmp.bin'
  gcc_cmd += ' -O{0}'.format(cfg['opt_level'])
  for flag in GCC_FLAGS:
    if cfg[flag] == 'on':
      gcc_cmd += ' -f{0}'.format(flag)
    elif cfg[flag] == 'off':
      gcc_cmd += ' -fno-{0}'.format(flag)
  for param, min, max in GCC_PARAMS:
    gcc_cmd += ' --param {0}={1}'.format(
      param, cfg[param])

  compile_result = self.call_program(gcc_cmd)
  assert compile_result['returncode'] == 0
  run_result = self.call_program('./tmp.bin')
  assert run_result['returncode'] == 0
  return Result(time=run_result['time'])
```



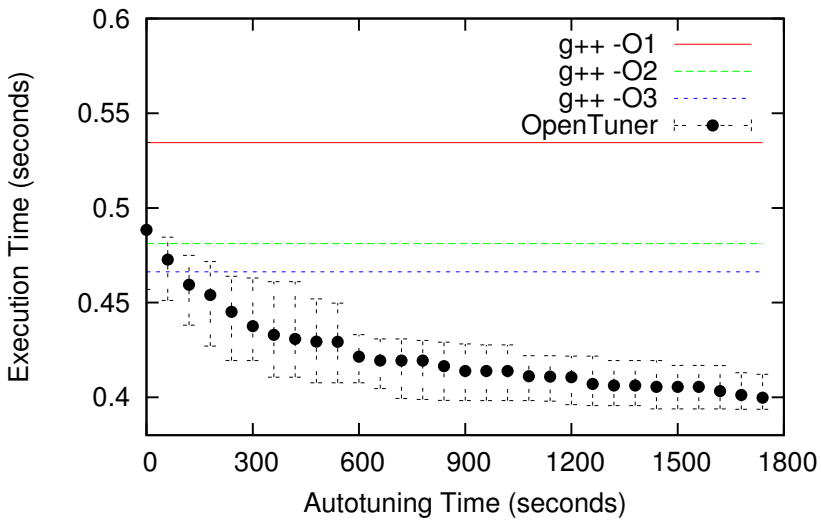
(a) fft.c



(b) matrixmultiply.cpp



(c) raytracer.cpp



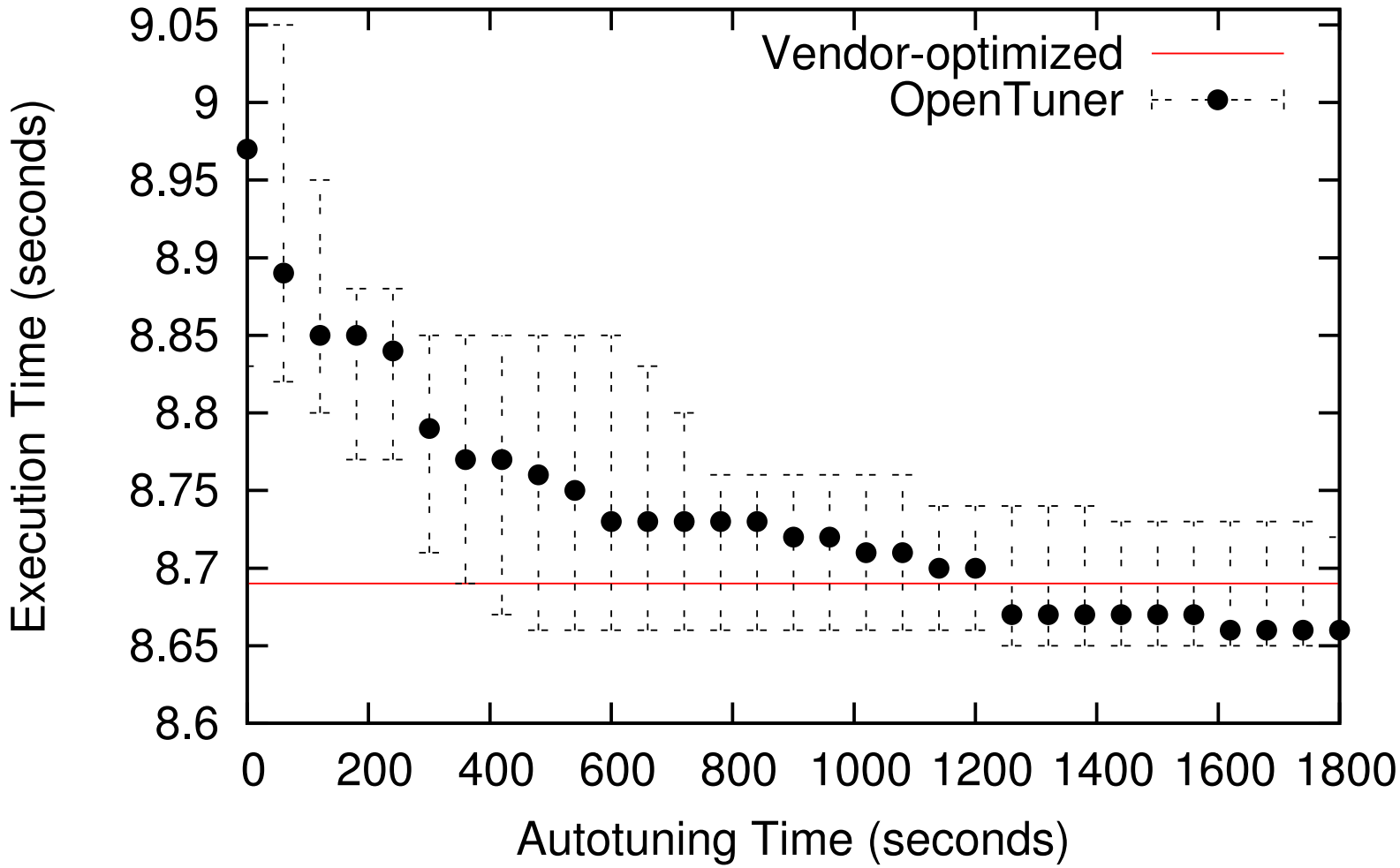
(d) tsp_ga.cpp

Konkrete Anwendungen der Entwickler

- vergleichbare Anwendung: Halide, PetaBricks, Stencil
- TODO: Bezug nehmen oder neue 1 weitere beschreiben?

High Performance Linpack Benchmark (HPL)

- Bestimmen der `floating point performance` von Rechnern
- Kriterium für Wahl der Top 500 Supercomputer
- HPL misst Geschwindigkeit für Lösen eines großen LGS
- HPL besitzt eingebauten Autotuner (vollst. Suche)
- ca. 15 versch. Parameter (u. a. Matrix-Blockgröße)
- größter Performance-Einfluss: Blockgröße (`IntegerParameter`)
- OpenTuner im Vergleich zu HPL-Implementierung von Intel


Vendor-optimized
OpenTuner
Execution Time (seconds)
Autotuning Time (seconds)
9.05
9
8.95
8.9
8.85
8.8
8.75
8.7
8.65
8.6
0
200
400
600
800
1000
1200
1400
1600
1800

Super Mario

- Abschließen des ersten Levels durch Sequenz von Button-Eingaben
-

- abc abc abc