Evaluating Design Tradeoffs in Numeric Static Analysis for Java

Real leader of this effort!

Shiyi Wei, Piotr Mardziel, Andrew Ruef, Jeffrey S. Foster, and Michael Hicks



Work done while all authors at

University of Texas at Dallas,

Carnegie Mellon,

2011 Lab for Programming Languages Research at the University of Maryland

Numeric Static Analysis (for Java)

- Prove numeric properties about programs, to show
 - array indexes in bounds,
 - division never by zero,
 - running times do not exceed limit, etc.
 - -Previous work: Use numeric analysis to compute running times, and see if they depend on a secret's value (PLDI'17)
 - Often framed as abstract interpretation
- Challenge: Analyzer that works for full Java
 - Aim to be sound(y), precise, scalable

The Problem

- •No prior analysis/paper soundily models
 - the heap,
 - method calls, and
 - numeric domains
 - In a way that scales, and requires few/no annotations
- Many design choices in building such an analysis
 - Lots of folklore about these choices

-Possibly overfit to evaluation target programs?

- Most papers consider only single choices
 - -Not their interactions

Prior Work

- •PAGAI: numeric analysis, but no heap
- •Fu (2014) considers both, but no method calls
- Ferrara; Magill; McCloskey et al; Chang and Rival:
 Sophisticated numeric+heap invariants but
 - not scalable, requires annotations, may be missing language features (e.g., method calls or loops)
- Little work analyzes interacting tradeoffs
 - ASTREE has many ideas, but not considered systematically (just final result indicated in paper)
 - Some one-offs; eg., polyhedra vs. intervals

Our work

- •We built a Java analysis tool that has multiple implementations of five analysis components
 - numeric domain (2)
 - heap abstraction (3)
 - abstract object representation (3) 162 configurations
 - interprocedural analysis order (3)
 - level of context sensitivity (3)
- Ran each on DaCapo, measured performance and precision on proving array indexes in bounds
- •Analyzed data to understand tradeoffs

Analysis Methodology

- Multiple linear (and logistic) regression
 - Predict precision and performance (dependent vars) as a linear function of analysis options (independent vars)

-Programs are indep. vars, to isolate effects of size, complexity, ...

- -Also considered two-way interactions of options as indep. vars
- Assuming a good model fit, doing this helps understand the complex space of possible options
- Visualization plotting precision vs. performance, per program, for all options
 - To see local variations, sanity-check trends

Results

- •Results clarify tradeoffs among analysis options
 - Yes, convex polyhedra are more precise, but slower, than intervals; but our results say how much,
 - and how these tradeoffs relate to that of varying the heap model (e.g., summary objects vs. access paths)
- General: Need more empirical work
 - Research often focuses on novelty, and math
 - Need as much or more work on engineering and measurement
 - Our methodology can help show the way

- Paper at https://arxiv.org/abs/1802.08927
 - Longer version of paper from ESOP'18
- •Code at https://github.com/plum-umd/JANA

Config. Option	Setting	Description
Numeric domain	INT	Intervals
(ND)	POL	Polyhedra
Hoop obstraction	SO	Only summary objects
$(\mathbf{H}\mathbf{\Lambda})$	AP	Only access paths
	AP+SO	Both access paths and summary objects
Abstract object	ALLO	Alloc-site abstraction
representation	CLAS	Class-based abstraction
(OR)	SMUS	Alloc-site except Strings
Inter-procedural	TD	Top-down
analysis order	BU	Bottom-up
(AO)	TD+BU	Hybrid top-down and bottom-up
Contort	CI	Context-insensitive
context	1CFA	1-CFA
$\int SETISTUTVITY (CS)$	1TYP	Type-sensitive

non-relational numeric domain (e.g., intervals)

$$[x == 1] \xrightarrow{x} = 1;$$

$$y = x;$$

$$[y == 1] \xrightarrow{x} = [console input];$$

$$[x == T] \xrightarrow{y} = x;$$

$$[y == T] \xrightarrow{x} = [console input];$$

$$y = x;$$

$$[y == T] \xrightarrow{x} = [console input];$$

relational numeric domain (e.g., convex polyhedra)



Tradeoff I: Numeric Domain



Adding the Heap: Points-to Analysis

- •Use points-to analysis to reason about aliasing
 - Doing x.f=3 could affect y.f if x and y are aliases
- Compute **Pt**: *Path* \rightarrow *P*(*Loc*) where
 - Path: a variable z or access path x.f and
 - Loc o represents an abstract location

-The name o represents one or more run-time values

- Hence $Pt(p) \cap Pt(q) \neq \emptyset$ implies p and q may alias

•Using **Pt**, we define a heap abstraction to integrate with our numeric domain

Access Paths [De and De'Souza ECOOP'12, Wei and Ryder ECOOP'14]

- Treat x. f as a numeric "variable"
 - Allow strong updates to it
 - But weakly update paths y.f where y may alias x

Strong update:

Weak update:

replace current abstraction

extend/overapprox current abstraction

$$\begin{bmatrix} x == \bot \end{bmatrix} \longrightarrow x = I; \qquad \begin{bmatrix} x == \bot \end{bmatrix} \\ x = 2; \qquad \begin{bmatrix} x == 1 \end{bmatrix} \\ x = 2; \qquad \begin{bmatrix} 1 \le x \le 2 \end{bmatrix}$$
 14

Access Paths Example





Summary Objects [Gopan et al, TACAS'04]

- Model o.f as a "summary" numeric object
 - When assigning to x.f, weakly update path o.f for all $o \in \mathbf{Pt}(x)$
- Harms and helps precision
 - No strong updates, hurts precision
 - But: Global nature of points-to analysis allows summary objects to account for effects across method calls
 - And makes expressible invariants like o.f < o.g</p>
- Harms performance: many objects in abstract state; weak updates slow (due to joins)

Summary Objects Example



Combination of the AP and SO also possible



Tradeoff II: Heap Abstraction



Tradeoff III: Abstract Object Representation (AOR)

Recall: **Pt**: Path
$$\rightarrow P(\text{Loc})$$
.

How should we represent Loc, i.e., abstract objects?

L1: Circle x = new Circle();
L2: Circle w = new Circle();



L1: Circle x = new Circle(); L2: Circle w = new Circle();



AOR Effect on Access Paths



AOR also affects summary object representation

22

Tradeoff III: Abstract Object Representation



- The smush-string option is a hybrid
 - all Java String instances use a single abstract object
 - other objects uses allocation-based approach

```
void foo() {
  Circle x = new Circle();
  Circle z = new Circle();
  x.radius = 1;
  z.radius = 2;
  int i = bar(x);
  int j = bar(z);
  assert (i + j == 5);
}
int bar(Circle p) {
  return p.radius + 1;
}
```

```
void foo() {
  L1: Circle x = new Circle();
  L2: Circle z = new Circle();
  x.radius = 1;
  z.radius = 2;
  L3: int i = bar(x);
  L4: int j = bar(z);
  assert (i + j == 5);
}
int bar(Circle p) {
  return p.radius + 1;
}
```



call-site sensitivity (ICFA)



```
void foo() {
  L1: Circle x = new Circle();
  L2: Circle z = new Circle();
  x.radius = 1;
  z.radius = 2;
  L3: int i = bar(x);
  L4: int j = bar(z);
  assert (i + j == 5);
```

int bar(Circle p) { return p.radius + 1;

}

}

Also affects precision of call graph and the way method summaries are stored

More Precise

Context Sensitivity vs. Type Sensitivity vs. Context Insensitivity



• Type-sensitive analysis distinguishes the function calls based on the type of the receiver

- **Top-down** interprocedural analysis
 - Analyzes each method starting at main, following the call graph down; carries forward access paths
- Precise analysis of each call
 - Boosted by context sensitivity if enabled
- •But may reanalyze same method many times
 - But: we cache a summary of a method's analysis, indexed by facts about its arguments
 - Reuse summary if subsequent calls' arguments' abstractions covered by those of previous analysis

Tradeoff IV: Inter-procedural Analysis Order



- Bottom-up analysis
 - Analyzes each method starting at leaves, generating a summary that can be used at each caller
- •No re-analysis, amenable to parallelization
 - Context sensitivity requires multiple summaries for each method (one per context)
- May lose precision depending on summary format
 - Convex polyhedra permit expressing relationships between arguments and return value; but not intervals
- •Complicated by use of summary objects
 - Harms performance and precision

Tradeoff IV: Inter-procedural Analysis Order

Tradeoff V: Inter-procedural Analysis Order



- Hybrid TD+BU option performs
 - bottom-up analysis for the Java library methods and
 - top-down analysis for the application code

Our tool: JANA (Java Numeric Analyzer)

https://github.com/plum-umd/JANA

- •WALA Java analysis framework
 - To parse bytecode to IR
 - To perform points-to analysis, CFG construction
- •Numeric domain backends
 - APRON for Intervals
 - ELINA for Convex Polyhedra
- 14K lines of Scala code

Experimental Evaluation

- RQI: performance
 - How does the configuration affect running time?
- RQ2: precision
 - How does the configuration affect precision?
- RQ3: tradeoffs
 - How does the configuration affect the precision/performance tradeoff?

Note: a configuration is one combination of the analysis configuration options.

Evaluation: Setup

- Analysis client: array index-out-of-bound analysis
- DaCapo benchmark suite
 - most popular Java benchmark

Program	Size (# IR	# Checks
	Instructions)	
antlr	55734	1526
bloat	150197	4621
chart	167621	7965
eclipse	18938	1043
fop	33243	1337
hsqldb	19497	1020
jython	127661	4232
luindex	69027	2764
lusearch	20242	1062
pmd	116422	4402
xalan	20315	1043

Evaluation: Setup

- Analysis client: array index-out-of-bound analysis
- DaCapo benchmark suite
 - most popular Java benchmark
 - Il real-world programs (some very large)
- 162 analysis configurations
 - vary each of the analysis options
 - 3 runs for each configuration on each program
- •Timeout: I hour per run
 - I week to run the whole thing on one machine

Setup Caveats

- •Results may change with different analysis client
 - Or benchmark suite
- •Three trials may be insufficient
 - Average variance ((max-min)/median) of all runs: 4.2%
 most runs differ by at most 4 mins (outlier: 32 mins, eclipse)
- Our implementation may have flaws
 - Though we used time-tested libraries and tools and tried to follow best practice

Evaluation: Data Analysis

- Multiple linear regression
 - dependent variables: performance and/or whether run times out (RQI) and precision (RQ2)
 - independent variables
 - -configuration options (and their two-way interactions)
 - -benchmark program: allow us roughly factor out programspecific sources of performance or precision gain/loss (e.g., size)
- •Used Akaike Information Criterion (AIC) to drop insignificant variables from the model
- Model R²: Performance: 0.72, precision: 0.98
 Good fit

- In performance analysis, timed out runs are counted as if taking I hour
 - Thus indicate a lower bound of the true cost
- •In precision analysis, timed out runs ignored
 - Too little data means lower statistical significance
- •Separate timeout analysis in paper
 - How likely are particular options to induce timeout?
 - Teaser: Using AP rather than SO reduces timeout likelihood by 40,000 times (!)



- Summary objects incur a significant slowdown
 - (AP is 37.6 minutes faster)

Option	$\mathbf{Setting}$	Est. (#)	CI	p-value
	TD	-	-	-
AO	BU	-1.98	[-6.3, 1.76]	0.336
	TD+BU	1.97	[-1.78, 6.87]	0.364
	AP+SO		-	-
HA	AP	-37.6	[-42.36, -32.84]	< 0.001
	SO	0.15	$[-4.60, \ 4.91]$	0.949
	1TYP	-	-	-
CS	CI	-7.09	[-10.89, -3.28]	< 0.001
	1CFA	1.62	[-2.19, 5.42]	0.405
	ALLO	-	-	-
OR	CLAS	-11.00	[-15.44, -6.56]	< 0.001
	SMUS	-7.15	[-11.59, -2.70]	0.002
ND	POL	-	-	-
	INT	-16.51	[-19.56, -13.46]	< 0.001

• The polyhedral domain is slow, but not as slow as summary objects

Option	$\mathbf{Setting}$	Est. (#)	CI	p-value
	TD	-	-	-
AO	BU	-1.98	[-6.3, 1.76]	0.336
	TD+BU	1.97	$[-1.78, \ 6.87]$	0.364
	AP+SO	-	-	-
HA	AP	-37.6	[-42.36, -32.84]	< 0.001
	SO	0.15	$[-4.60, \ 4.91]$	0.949
	1TYP	-	-	-
CS	CI	-7.09	[-10.89, -3.28]	< 0.001
	1CFA	1.62	[-2.19, 5.42]	0.405
	ALLO	-	-	-
OR	CLAS	-11.00	[-15.44, -6.56]	< 0.001
	SMUS	-7.15	[-11.59, -2.70]	0.002
ND	POL	-	-	-
	INT	(-16.51	[-19.56, -13.46]	< 0.001

 Heavyweight CS and OR settings hurt performance, particularly when using summary objects

[Option	$\mathbf{Setting}$	Est. (#)	\mathbf{CI}	p-value
		TD	-	-	-
	AO	BU	-1.98	[-6.3, 1.76]	0.336
		TD+BU	1.97	$[-1.78, \ 6.87]$	0.364
		AP+SO	-	-	-
	HA	AP	-37.6	[-42.36, -32.84]	< 0.001
		SO	0.15	$[-4.60, \ 4.91]$	0.949
		1TYP	-	-	-
	CS	CI	(-7.09) [-10.89, -3.28]	< 0.001
		1CFA	1.62	[-2.19, 5.42]	0.405
		ALLO	-	-	-
	OR	CLAS	-11.00	[-15.44, -6.56]	< 0.001
		SMUS	-7.15	[-11.59, -2.70]	0.002
	ND	POL	-	-	-
		INT	-16.51	[-19.56, -13.46]	< 0.001
		AP+SO:ALLO	-	-	-
		AP:CLAS	9.55	[5.37, 13.71]	< 0.001
ower is t	HA:OR	AP:SMUS	6.25	[2.08, 10.42]	< 0.001
		SO:SMUS	0.07	[-4.09, 4.24]	0.973
		SOCLAS	-0.43	$\begin{bmatrix} -4 59 3 73 \end{bmatrix}$	0.839

46

 Bottom-up analysis does not provide a performance advantage

Option	$\mathbf{Setting}$	Est. (#)	CI	p-value
	TD	-	-	
AO	BU	-1.98	[-6.3, 1.76]	0.336
	TD+BU	1.97	$[-1.78, \ 6.87]$	0.364
	AP+SO	-	-	-
HA	AP	-37.6	[-42.36, -32.84]	< 0.001
	SO	0.15	$[-4.60, \ 4.91]$	0.949
	1TYP	-	-	-
CS	CI	-7.09	[-10.89, -3.28]	< 0.001
	1CFA	1.62	[-2.19, 5.42]	0.405
	ALLO	-	-	-
OR	CLAS	-11.00	[-15.44, -6.56]	< 0.001
	SMUS	-7.15	[-11.59, -2.70]	0.002
ND	POL	-	-	-
	INT	-16.51	[-19.56, -13.46]	< 0.001

	# indexes proved in bounds				
		Ļ			
Option	Setting	Est. (#)	\mathbf{CI}	p-value	
	TD	-	-	_	
AO	TD+BU	-134.22	[-184.93, -83.50]	< 0.001	
	BU	-129.98	[-180.24, -79.73]	< 0.001	
	AP+SO	-	-	-	
HA	SO	-94.46	[-166.79, -22.13]	0.011	
	AP	-5.24	[-66.47, 55.99]	0.866	
	ALLO	-	_	-	
OR	CLAS	-90.15	[-138.80, -41.5]	< 0.001	
	SMUS	35.47	[-14.72, 85.67]	0.166	
	POL	-	-	-	
	INT	5.11	[-28.77, 38.99]	0.767	

(Higher is better)

Access paths are critical to precision

Option	Setting	Est. (#)	\mathbf{CI}	p-value
	TD	-	-	-
AO	TD+BU	-134.22	[-184.93, -83.50]	< 0.001
	BU	-129.98	[-180.24, -79.73]	< 0.001
	AP+SO	-	-	-
HA	SO	-94.46	[-166.79, -22.13]	0.011
	AP	-5.24	[-66.47, 55.99]	0.866
	ALLO	-	-	-
OR	CLAS	-90.15	[-138.80, -41.5]	< 0.001
	SMUS	35.47	[-14.72, 85.67]	0.166
	POL	-	-	-
	INT	5.11	[-28.77, 38.99]	0.767

(Higher is better)

 Bottom-up analysis harms precision overall, especially for SO (only)

	Option	Setting	Est. (#)	\mathbf{CI}	p-value
		TD	-	-	-
	AO	TD+BU	-134.22	[-184.93, -83.50]	< 0.001
		BU	-129.98	[-180.24, -79.73]	< 0.001
		AP+SO	-	-	-
	HA	SO	-94.46	[-166.79, -22.13]	0.011
		AP	-5.24	[-66.47, 55.99]	0.866
		ALLO	-	-	-
	OR	CLAS	-90.15	[-138.80, -41.5]	< 0.001
		SMUS	35.47	[-14.72, 85.67]	0.166
	ND	POL	-	-	-
		INT	5.11	[-28.77, 38.99]	0.767
		TD:AP+SO	-	-	-
		BU:SO	-686.79	[-741.82, -631.76]	< 0.001
	AO:HA	TD+BU:SO	-630.99	[-687.41, -574.56]	< 0.001
(Higher is		TD+BU:AP	63.59	[14.71, 112.47]	0.011
		BU:AP	58.92	[11.75, 106.1]	0.014

50

•The relational domain improves precision some

	Option	Setting	Est. $(#)$	\mathbf{CI}	p-value
		TD	-	-	-
	AO	TD+BU	-134.22	[-184.93, -83.50]	< 0.001
		BU	-129.98	[-180.24, -79.73]	< 0.001
		AP+SO	-	-	-
	HA	SO	-94.46	[-166.79, -22.13]	0.011
		AP	-5.24	[-66.47, 55.99]	0.866
	OR	ALLO	-	-	-
		CLAS	-90.15	[-138.80, -41.5]	< 0.001
		SMUS	35.47	[-14.72, 85.67]	0.166
	ND	POL	-	-	-
		INT	5.11	[-28.77, 38.99]	0.767
		AP+SO:POL	-		-
	HA:ND	AP:INT	-58.87	[-99.39, -18.35]	0.004
		SO:INT	-61.96	[-109.08, -14.84]	0.01
(Higher is	bette	er)			

•More precise OR improves precision, but not CS (not in model)

Option	Setting	Est. (#)	\mathbf{CI}	p-value
	TD	-	-	-
AO	TD+BU	-134.22	[-184.93, -83.50]	< 0.001
	BU	-129.98	[-180.24, -79.73]	< 0.001
	AP+SO	-	-	-
HA	SO	-94.46	[-166.79, -22.13]	0.011
	AP	-5.24	[-66.47, 55.99]	0.866
	ALLO	-	-	-
OR	CLAS	-90.15	[-138.80, -41.5]	< 0.001
	SMUS	35.47	[-14.72, 85.67]	0.166
	POL	-	-	-
	INT	5.11	[-28.77, 38.99]	0.767

RQ3: Tradeoffs: Best and Worst

		#	Best I	Performan	ce	Bes	st Precision	ı
Prog	Size	Checks	$\left \operatorname{Time}(\min)\right $	# Checks	Percent	Time(min)	# Checks	Percent
			BU-AP	-CI-CLAS-	INT	TD-AP+SO	O-1TYP-CL	AS-INT
antlr	55734	1526	0.6	1176	77.1%	18.5	1306	85.6%
			BU-AP	-CI-CLAS-	INT	TD-AP-1	TYP-SMUS	S-POL
bloat	150197	4621	4.0	2538	54.9%	17.2	2795	60.5%
			BU-AP	-CI-CLAS-	INT	TD-AP-1	TYP-SMU	S-INT
chart	167621	7965	3.3	5593	70.2%	7.7	5654	71.0%
			BU-AP	-CI-ALLO-	INT	TD-AP+SC)-1TYP-SM	IUS-POL
eclipse	18938	1043	0.2	896	85.9%	3.3	977	93.7%
			BU-AP	-CI-CLAS-	INT	TD-AP+SC	D-1CFA-SM	IUS-INT
fop	33243	1337	0.4	998	74.6%	2.6	1137	85.0%
			BU-AP-	-CI-SMUS-	INT	TD-AP+S	SO-CI-SMU	JS-INT
hsqldb	19497	1020	0.3	911	89.3%	1.4	975	95.6%
			BU-AP-	-CI-SMUS-	INT	TD-AP-1	ICFA-CLAS	S-POL
jython	127661	4232	1.3	2667	63.0%	33.6	2919	69.0%
			BU-AP-	-CI-SMUS-	INT	TD-AP+SC	D-1TYP-AL	LO-INT
luindex	69027	2764	1.8	1682	60.9%	46.8	2015	72.9%
			BU-AP	-CI-CLAS-	INT	TD-AP+SC	D-1CFA-AL	LO-POL
lusearch	20242	1062	0.2	912	85.9%	54.2	979	92.2%
			BU-AP	-CI-CLAS-	INT	TD-AP+	SO-CI-CLA	S-INT
pmd	116422	4402	1.7	3153	71.6%	49.5	3301	75.0%
			BU-AP	-CI-CLAS-	INT	TD-AP+SC	D-1CFA-SM	US-POL
xalan	20315	1043	0.2	912	87.4%	3.8	981	94.1%

RQ3: Tradeoffs: Best and Worst

		#	Best Performance		Best Precision			
Prog	Size	Checks	$ \text{Time}(\min) #$	Checks	Percent	$\operatorname{Time}(\min) _{i}$	# Checks	Percent
			BU-AR-C	I-CLAS-I	NT	TD-AP+SO	-1TYP-CL	AS-INT
antlr	55734	1526	0.6	1176	77.1%	18.5	1306	85.6%
			BU-AP-C	I-CLAS-I	NT	TD-AP-1	TYP-SMUS	S-POL
bloat	150197	4621	4.0	2538	54.9%	17.2	2795	60.5%
			BU-AP-C	I-CLAS-I	NT	TD-AP-1	TYP-SMU	S-INT
chart	167621	7965	3.3	5593	70.2%	7.7	5654	71.0%
			BU-AP-C	I-ALLO-I	NT	TD-AP+SO	-1TYP-SM	US-POL
eclipse	18938	1043	0.2	896	85.9%	3.3	977	93.7%
			BU-AP-C	I-CLAS-I	NT	TD-AP+SO	-1CFA-SM	US-INT
fop	33243	1337	0.4	998	74.6%	2.6	1137	85.0%
			BU-AP-C	I-SMUS-I	NT	TD-AP+S	0-CI-SMU	S-INT
hsqldb	19497	1020	0.3	911	89.3%	1.4	975	95.6%
			BU-AP-C	I-SMUS-I	NT	TD-AP-1	CFA-CLAS	-POL
jython	127661	4232	1.3	2667	63.0%	33.6	2919	69.0%
			BU-AP-C	I-SMUS-I	NT	TD-AP+SO	-1TYP-AL	LO-INT
luindex	69027	2764	1.8	1682	60.9%	46.8	2015	72.9%
			BU-AP-C	CLAS-I	NT	TD-AP+SO	1CFA-ALI	_O-POL
lusearch	20242	1062	0.2	912	85.9%	54.2	979	92.2%
			BU-AP-Q	I-CLAS-I	NT	TD-AP+	O-CI-CLA	S-INT
pmd	116422	4402	1.7	3153	71.6%	49.5	3301	75.0%
			BU-AP-C	I-CLAS-I	NT	TD-AP+90	-1CFA-SM	US-POL
xalan	20315	1043	0.2	912	87.4%	3.8	981	94.1%

RQ3: Tradeoffs: Best and Worst

		#	Best Performance		Best Precision			
Prog	Size	Checks	$\operatorname{Time}(\min)$	# Checks	Percent	$\left \operatorname{Time}(\min)\right $	# Checks	Percent
			BU-AP	-CI-CLAS-	INT	TD-AP+SC)-1TYP-CL	AS-INT
antlr	55734	1526	0.6	1176	77.1%	18.5	1306	85.6%
			BU-AP	-CI-CLAS-	INT	TD-AP-1	TYP-SMUS	S-POL
bloat	150197	4621	4.0	2538	54.9%	17.2	2795	60.5%
			BU-AP	-CI-CLAS-	INT	TD-AP-1	TYP-SMU	S-INT
chart	167621	7965	3.3	5593	70.2%	7.7	5654	71.0%
			BU-AP	-CI-ALLO-	INT	TD-AP+SO	-1TYP-SM	US-POL
eclipse	18938	1043	0.2	896	85.9%	3.3	977	93.7%
			BU-AP	-CI-CLAS-	INT	TD-AP+SC	D-1CFA-SM	US-INT
fop	33243	1337	0.4	998	74.6%	2.6	1137	85.0%
			BU-AP-	-CI-SMUS-	INT	TD-AP+S	SO-CI-SMU	IS-INT
hsqldb	19497	1020	0.3	911	89.3%	> $<$ 1.4	975	95.6%
			BU-AP-	-CI-SMUS-	INT	TD-AP-1	CFA-CLAS	-POL
jython	127661	4232	1.3	2667	63.0%	33.6	2919	69.0%
			BU-AP-	CI-SMUS-	INT	TD-AP+SC)-1TYP-AL	LO-INT
luindex	69027	2764	1.8	1682	60.9%	\triangleright 46.8	2015	72.9%
			BU-AP	-CI-CLAS-	NT	TD-AP+SC)-1CFA-AL	LO-POL
lusearch	20242	1062	0.2	912	85.9%	54.2	979	92.2%
			BU-AP	-CI-CLAS-	INT	TD-AP+S	SO-CI-CLA	S-INT
pmd	116422	4402	1.7	3153	71.6%	49.5	3301	75.0%
			BU-AP	-CI-CLAS-	INT	TD-AP+SC	-1CFA-SM	US-POL
xalan	20315	1043	0.2	912	87.4%	3.8	981	94.1%

RQ3: Tradeoffs



(More graphs in paper than I can show now)

RQ3: Tradeoffs

 Access paths improve precision with good performance; summary objects slow and imprecise



RQ3: Tradeoffs \circ BU \triangle TD + TD+BU

• Top-down analysis is preferred: bottom-up is less precise and does little to improve performance



RQ3: Tradeoffs - Interactions

• AP&POLY generally better than AP&INT



RQ3: Tradeoffs - Interactions

But AP+SO&INT better than AP+SO&POLY (fewer timeouts)



- Access paths are always a good idea
 - Add significant precision at little performance cost
- •Summary objects + access paths add precision
 - But adds significant performance overhead, often resulting in timeouts
- Polyhedra improve precision
 - But time out with AP+SO abstraction

-Intervals and AP+SO would work better

- Top-down more than precise bottom up
 - And loses little in performance

Methodology Summary

- Static analysis tools are complex, with many interacting options
- Need experimental work to understand tradeoffs
- •Our work is a template for future work
 - Measure effects of various options on good benchmark
 - Use MLR to understand impact of options, generally
 - And visualization to see local effects

Empirical Evaluations: Advice

- •SIGPLAN Guidelines for Empirical Evaluations
- Set up as a checklist to help design a good

evaluation



		SIGPLAN Empirical	Eval	uation	Checklist
2	¢	Explicit Claims Colms mad be explicit in order for the reader to essens whether the empirical evaluation supports them. Course aroutd aim to race not just what is extremed but here.	Meters		Direct or Appropriate Proop Metric If the meat velocity metric to not icr connect bein measured directly. The proop metric used measures in well particle. The overside, antiduction to code measure to rearise appropriate proop for actual and to end participation or a merge somewhyles.
Clearly District Claim Example from Press		Appropriately-Booped Daines The such of dates should follow from the arithmes pro- vided. Coardianting is often the assessments of inde- casts enterco.e.g., cheming levels of all Joint, but en- ualing only a platic subset or paining levels provide facts water, but evaluating any incorrection, emulation.	Relevant B		Measures. Millingonant Effects The cests and benefits of a technique may be multi-located. All tasks attoch to considered, both cases and benefits, and obely reduced. For example, complex technications may gread apprograms at the cost of theologicy increasing campor time.
	1	Advecentariges Cantanteres A paper shauld admonistrate in Enterfances to please the server of reacitie in-constant. Training no-initiations all all or only-categorial once while arbiting the main elevant terms, may-mailed the reader to drawing too-attong-conclusions.			Sufficient Information in Report Experiments should be described in sufficient detail to be repeated in Algorithmic Controlling Island valued should be included, an well as id and/or sufficient of adhesis, and full details of hardware platforms.
sparison of Proton	T	Appropriate likewither for Comparison An empirical evaluation of a contribution that improves upon the state-of-the-art alread available fragment as appropriate based of based comparison are a nonlocalizable.	mental Design		Researchite Pattern The sentiation should be on a patient that is necessarily be said to near-the olders. For example, a size that the basis is patienteness on notific platform alread net have an evolution performed exclusively on server.
Buitable Com Françai 3	<u>î</u>	Fair Comparison Concentrations to a comparing system should not unlikily dis- advantage that system. Nor existingle, steady, the compared systemic works like compiled with the same compiler and op- timization legal.	SCiter Experi- transfer the Par	4	Explores Key Design Hexanetiers New parameters induction exclosed over a surgerio evalu- ate sensitivity to their settings. Examples induct the size of the image when evaluating a foracity aptimization. All as- panded system-configurations (set), non-secting-to- set approximation over the section of the size of setting-orientation evaluating a foracity approximation. All as- panded system-configurations (setting or section) when over-
Choice	2	Appropriete Softe Evaluations should be conducted using the appropriate en- tablehed tendenaries where these ends. Exabilities using should be wanted in the designed for cardinality for assigning paneler performance.	Apropriate and	8 Q.	Open Loop in Workload Sevender Land perentam to tobial themselor-oriented systems should not be pated by the rate at which the seriem re- spanse. Head, the code pervessor should be core top; pervesting work independent of the performance of the sys- lem under test. Size (Biofreeder et al., 1998)
Principled Benchmark C Earthe Benchmark		Non-Granieri Bulaciji Avefilieri Barretiras er ostatished bereferasti sulla dass nil erist. A retarale etavid be providel for the selection of hene- gravn bereferantis er subsetling estatished bereferanti fulles.		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Gross Multivelies Means Readed When a system airs to be general hid van developed by training on an deale consideration of spacific avamples, it is assential had the revealation explicitly perform reso- volciation, so that the system is evel-and on tass defining from the training set.
		Applications, Rod (Acad) Harvaris A claim that is operant benefits around applications should be based on such applications dheady, and not only or micro-learned periods can be sated and appropriate, in a breader analysis(r)		34 A A A A A	Comprehensive Burnmary Results Appropriate visitable shauld be used to characterize fre-Lift range of results, not just the med tearnetite values, which may be automatic. For exercise, I is not appropriate to sum- matus speedups of 4%, 1%, and 40% as light 49%.
43	•	Sufficient Hamber of Trais In modern systems, which have non-deterministic perfor- ments, a small-number of trais long, is single time me- anements (name many non-an engine). Similary, make in- ale may be reached to get the system related trais (e.g., time a smally state that available were up official).	testion of Ferry	1.	Asso include Zero A surveyed graph (with an axis not including perc) can ex- signment the importance of a therees. While sporting in to the incomenting range of an axis can serve three and expective, here is a significant risk that the is makeating pagestally 7.2 is not introducing size that the sale is the range.
Adequate Data Araby Earth feet Press		Appropriate fourneary Statistics There are many summary statistics, and each presents an accurate view of a dataset any under appopriate circu- stances. The exempts, the generative mean structed only be used when company wases with different tanges, and the fourneous mean when company stats. When distributions from suffers, a nature intractive presented.	contacto Pressent		Ratics Plotted Correctly When ratios (e.g., securitized) are pictified an one graph, the size-off-the bars must be interchyparthmically proportion to the charge. For exempts, 20 and 64 eer rectiprophil, but their inter distance-them 1.2 dates not whend that. This mainted up afted can be provided either by using a topscale on by contracting to the provided either by using a topscale on by contracting to the provided either by using a topscale
	ii ii veer	Report Data Distribution Reporting just a measure of central tendency (e.g., a reser or method lab is replace the anteni of any new determinion. A measure of variability cap, variance, and deviation, submitted and/or confidence intervels heb to un- detated the distribution of the data.	Ann	<u></u>	Appropriate Level of Precision The number of significant digits should reliest the preci- sion of the experiment, Repeting improvements of '00.00C when the experimental area to i an example of rel- strict procision, misleading the reviewer's underdoming of the eignificance of the INEL
P	E Mater (News sile	plan or of Measurer Materia Biology in /	Are this E.C.	Berger, S. M. Blue	Kharn, W. Hauseirth, and M. Hoka for the ACM SIGPLANED

http://sigplan.org/Resources/EmpiricalEvaluation/