

Extending over-approximating models: an introduction to mixed models and separation logic

David Schmidt

Kansas State University
(visiting École Polytechnique)

`www.cis.ksu.edu/~schmidt`

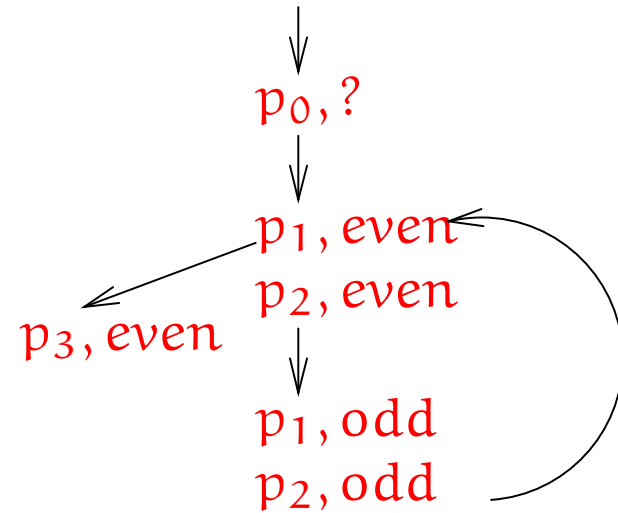
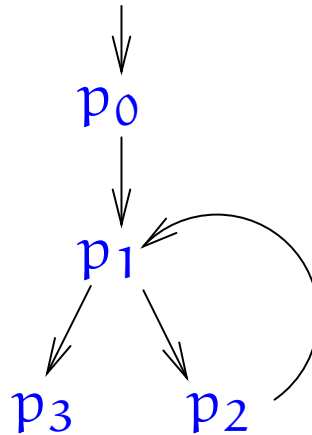
Outline

1. Review *over-approximating* relational models of control and data
2. Note that we can validate only *universal* properties on such models
3. Extend the models with *under-approximating relations* so that we can validate existentials. This gives us a *mixed model*.
4. Introduce modular validation, with an implicit existential, via *separation logic*
5. Note that the above approaches are *branching-time*, and that there is an equally interesting *linear-time* approach to validation

Models extracted from working hardware/software are *overapproximations*

Here is a program, its control-flow graph, and its polarity abstract interpretation:

p_0 : $x = 2$;
 p_1 : **while** $x \neq 0$ {
 p_2 : $x = x - 1$; }
 p_3 : *exit*



The program's concrete execution trace,

$p_0, ? \rightarrow p_1, 2 \rightarrow p_2, 2 \rightarrow p_1, 1 \rightarrow p_2, 1 \rightarrow p_1, 0 \rightarrow p_3, 0$

is modelled by abstract traces embedded in each of the two models.

The models use smaller state spaces but coarser transition relations.

A labelled Kripke transition system abstracts a process

$\langle \Sigma, \{\tau_\ell \subseteq \Sigma \times \Sigma \mid \ell \in \text{Label}\}, \mathcal{I}_\Sigma : \Sigma \rightarrow \mathcal{P}(\text{Atom}) \rangle$

$$\Sigma = \{s0, s1, s2\}$$

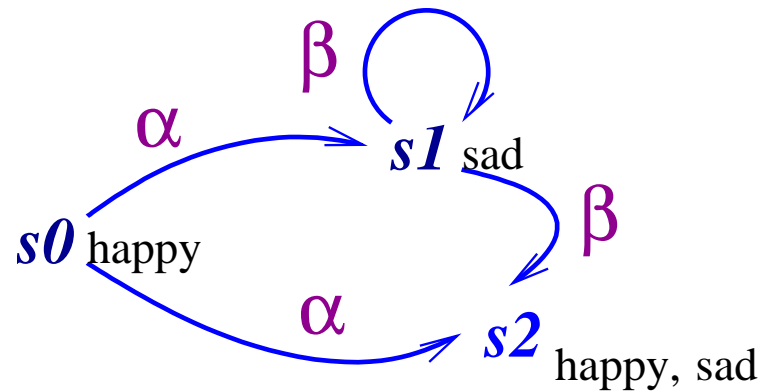
$$\tau_\alpha = \{(s0, s1), (s0, s2)\}$$

$$\tau_\beta = \{(s1, s1), (s1, s2)\}$$

$$\mathcal{I}_\Sigma(s0) = \{\text{happy}\}$$

$$\mathcal{I}_\Sigma(s1) = \{\text{sad}\}$$

$$\mathcal{I}_\Sigma(s2) = \{\text{happy, sad}\}$$



There are two transition relations, τ_α and τ_β .

A Kripke system might be approximated further:

$$A = \{\perp, a0, a12, \top\}$$

$$\alpha\{\} = \perp$$

$$\alpha\{c0\} = a0$$

$$\alpha\{c1\} = a12 = \alpha\{c2\} = \alpha\{c1, c2\}$$

$$\alpha S = \top, \text{ otherwise}$$

$$\gamma(\perp) = \{\}$$

$$\gamma(a0) = \{c0\}$$

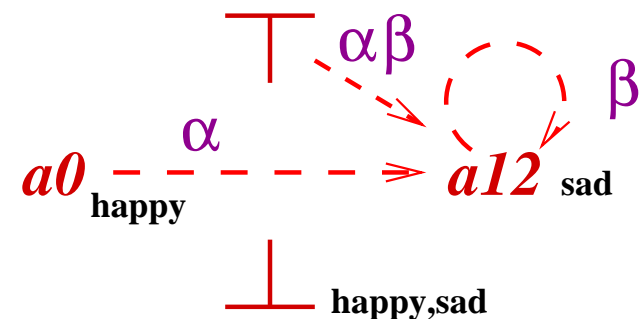
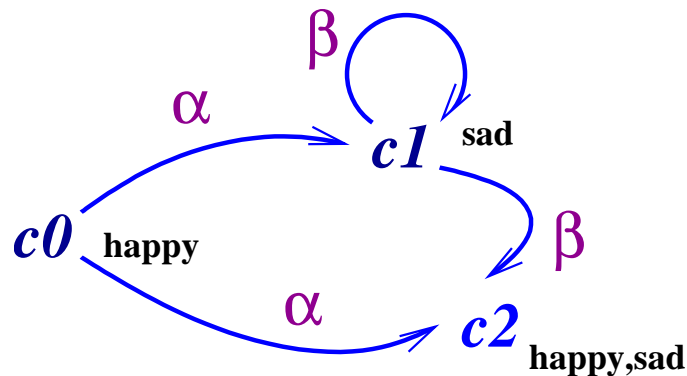
$$\gamma(a12) = \{c1, c2\}$$

$$\gamma(\top) = \{c0, c1, c2\}$$

$$\mathcal{I}_A(a) = \cap \{\mathcal{I}_C(c) \mid c \in \gamma(a)\}$$

α, γ form a *Galois connection*, induce the simulation relation, $\rho \subseteq \Sigma \times A$ — $c \rho a$ iff $c \in \gamma(a)$ — and define coarser α and β .

if $c \in \gamma(a)$
and $\downarrow \tau$ then $\downarrow \tau$
 $c' \in \gamma(a')$



Graph models apply to storage shapes

$$\Sigma = \{c0, c1, c2\}$$

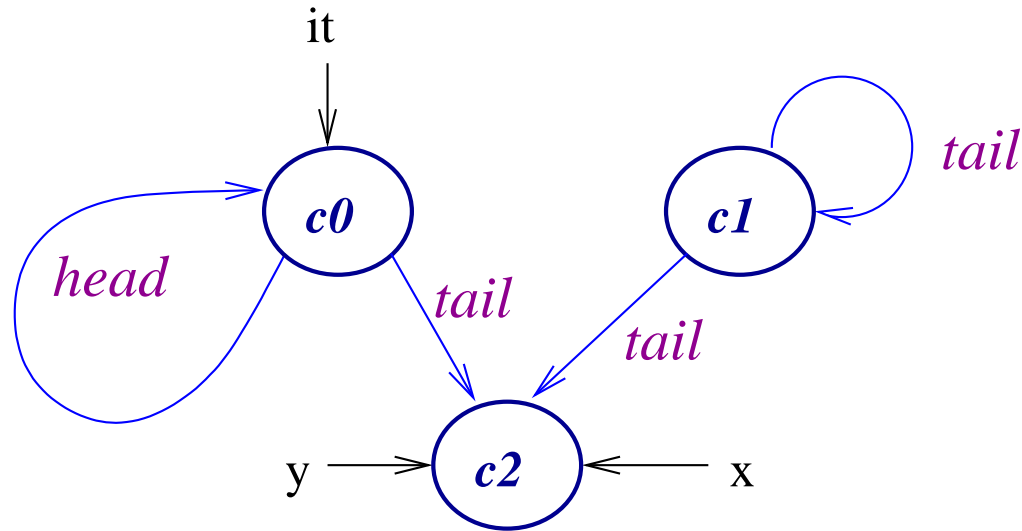
$$\tau_{\text{head}} = \{(c0, c0)\}$$

$$\tau_{\text{tail}} = \{(c0, c2), \\ (c1, c1), (c1, c2)\}$$

$$\mathcal{I}_{\Sigma}(c0) = \{\text{it}\}$$

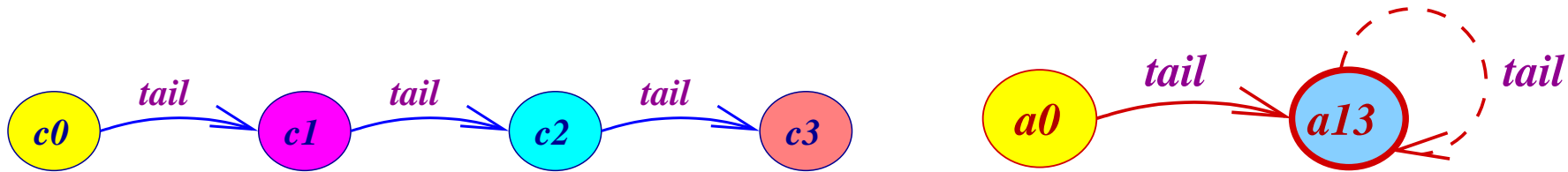
$$\mathcal{I}_{\Sigma}(c1) = \{\}$$

$$\mathcal{I}_{\Sigma}(c2) = \{x, y\}$$



Rather than states, the nodes now represent *cells*.

A linear list can be abstracted to its “shape graph”:



The noncircular list is modelled by a more compact representation that distinguishes only between the head and non-head nodes.

A data base can be abstracted by forgetting a dimension:

<i>surname</i>	<i>name</i>	<i>eye color</i>
Ricardo	Ricky	brown
Ricardo	Lucy	blue
Mertz	Fred	green

<i>surname</i>	<i>eye color</i>
Ricardo	brown \sqcup blue
Mertz	green

The abstraction produces a smaller element set with loss of precision.

Specifications/invariants are *underapproximations* of correct behaviours

An acceptable database must associate a unique eye color to each person:

$$\forall p : \text{Person}, \exists! c : \text{Color}, c \in p.\text{eyeColor}$$

An acceptable list must be noncircular:

$$\forall n \geq 0, (\forall i \in 0..n-1, a_{i+1} = a_i.\text{tail}) \wedge (\forall i, j \in 0..n, i \neq j \Rightarrow a_i \neq a_j)$$

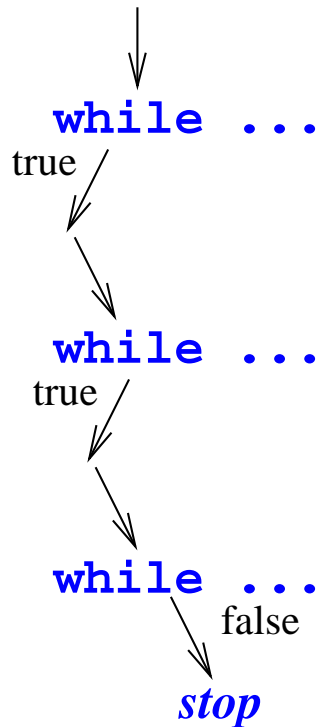
An acceptable program execution must terminate:

$$p_0 \models^\forall F(\text{at } p_3)$$

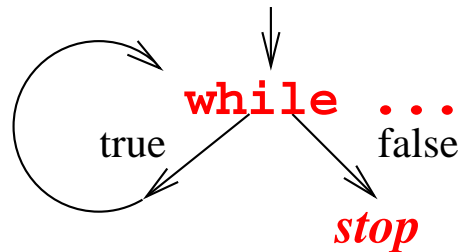
Alas, *none* of these properties can be proved for the overapproximating models we just saw!

Overapproximating models add extra transitions that generate nonexecuted paths. These paths reduce the properties we can prove on the overapproximating model (that also hold true for the concrete structure being modelled).

Concrete trace:



Overapproximation:



On the **overapproximation**, we cannot prove

\forall paths,

termination occurs

although it holds for the concrete semantics,

and we cannot **safely**

prove

\exists path that is infinite

because it does **not** hold for the concrete semantics

So, what can we verify?

Given the α, γ machinery, we can soundly validate universal properties of nodes, a , in overapproximating models, M :

$$\phi ::= p \mid \phi \wedge \phi \mid \phi \vee \phi \mid \forall R.\phi \mid \mu Z.\phi \mid \nu Z.\phi \mid Z$$

(Note: This is a fragment of *description logic*.)

When $a \models \phi$, then for all $c \in \gamma(a)$, we have $c \models \phi$.

$$a \models \phi \text{ iff } p \in \mathcal{I}_\Sigma(a)$$

$$a \models \phi_1 \wedge \phi_2 \text{ iff } a \models \phi_1 \text{ and } a \models \phi_2$$

$$a \models \forall R.\phi \text{ iff for all } a' \in a.R^\#, a' \models \phi \text{ (where } R^\# \text{ is sound w.r.t. } \alpha, \gamma)$$

$$a \models \mu Z.\phi \text{ iff there exists } i \geq 0 \text{ such that } a \models \phi_i, \text{ where } \begin{cases} \phi_0 = \text{false} \\ \phi_{i+1} = [\phi_i/Z]\phi \end{cases}$$

$$a \models \nu Z.\phi \text{ iff for all } i \geq 0, a \models \phi_i, \text{ where } \begin{cases} \phi_0 = \text{true} \\ \phi_{i+1} = [\phi_i/Z]\phi \end{cases}$$

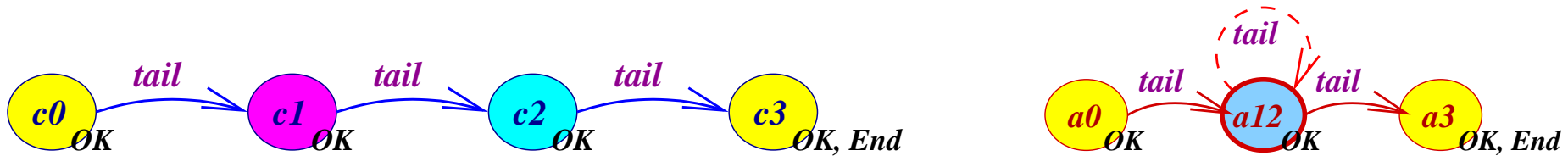
Note: The definitions of μ and ν assume that M is finite.

$a \models \phi_1 \wedge \phi_2$ iff $a \models \phi_1$ and $a \models \phi_2$

$a \models \forall R.\phi$ iff for all $a' \in a.R^\#$, $a' \models \phi$ (where $R^\#$ is sound w.r.t. α, γ)

$a \models \mu Z.\phi$ iff there exists $i \geq 0$ such that $a \models \phi_i$, where $\begin{cases} \phi_0 = \text{false} \\ \phi_{i+1} = [\phi_i/Z]\phi \end{cases}$

$a \models \nu Z.\phi$ iff for all $i \geq 0$, $a \models \phi_i$, where $\begin{cases} \phi_0 = \text{true} \\ \phi_{i+1} = [\phi_i/Z]\phi \end{cases}$



Let $\text{tail} \subseteq \text{Node} \times \text{Node}$ be the transition relation.

We can safely prove:

$a_0 \models \nu Z.\text{isOK} \wedge \forall \text{tail}.Z$: all nodes reached from a_0 (and so, c_0) are OK

$a_3 \models \mu Z.\text{atEnd} \vee \forall \text{tail}.Z$: all paths from a_3 (and so, c_3) terminate

We cannot prove: $a_0 \models \mu Z.\text{atEnd} \vee \forall \text{tail}.Z$ (but it holds for c_0)

We cannot safely prove:

$a_{12} \models \nu Z.\exists \text{next}.Z$: there is a cycle (because this fails for c_1 and c_2)

Why does this matter?

Program and data-structure models are almost always overapproximations. The specifications/invariants that we wish to validate often require a richer logic than the one just presented.

Can we do better?

1. Make the overapproximating model more detailed by adding *underapproximating relations*, dual to the ones we have. We check both universal and existential properties on the *mixed model*, and there are some specialization tricks (*generalized model checking*, *“focus” operations*, etc.) to improve precision.
2. Retain just the overapproximating model but add some key operators on relations that enrich the language of properties we might prove. This is a basis of *separation logic*.

1. Mixed models

An overapproximating model uses “maybe” relations, R^\sharp :

$(a, b) \in R^\sharp$, that is,

$$a \overset{R}{\dashv} \Rightarrow b$$

means “ a may be related to b ” in a concretization of this model.

(Precisely, $(u, v) \in R$ and $u \in \gamma(a)$ imply $(a, b) \in R^\sharp$ and $v \in \gamma(b)$.) When

$(a, b) \notin R^\sharp$, it signifies that *no* concretization of the model relates a concretization of a to b .

$$a \models \forall R. \phi \text{ iff for all } b \in a.R^\sharp, b \models \phi$$

The dual is a “must” relation: $(a, b) \in R^b$:

$$a \overset{R}{\longrightarrow} b$$

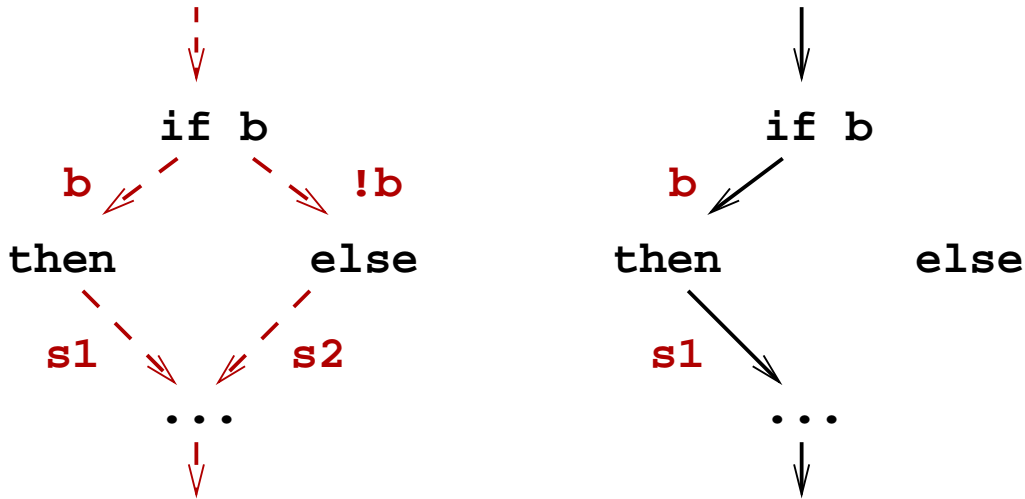
means “ a must be related to b ” in every concretization of this model.

(Precisely, $(a, b) \in R^b$ and $u \in \gamma(a)$ imply $(u, v) \in R$ and $v \in \gamma(b)$.)

$$a \models \exists R. \phi \text{ iff there exists } b \in a.R^b, b \models \phi$$

What does a “maybe” transition denote?

what may *possibly* execute:



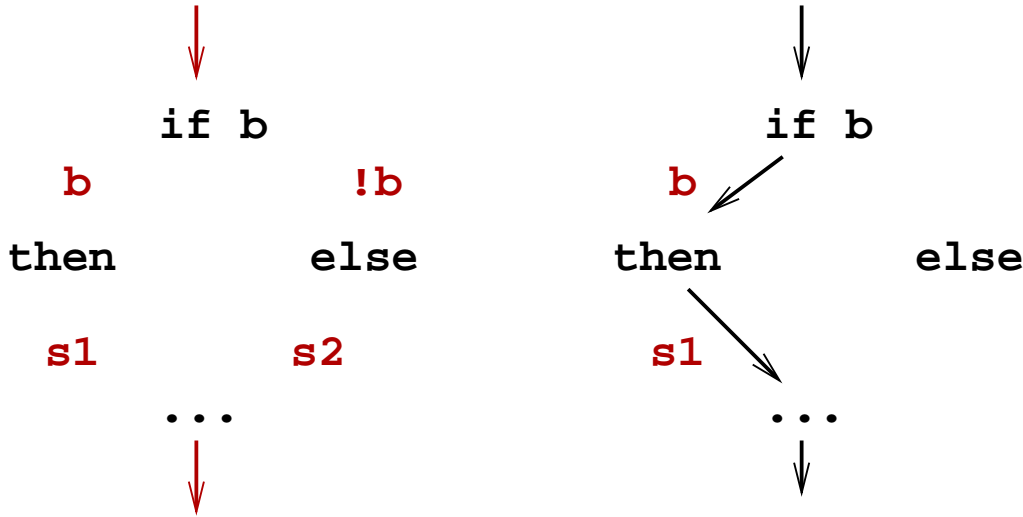
what may *possibly* be pointed:



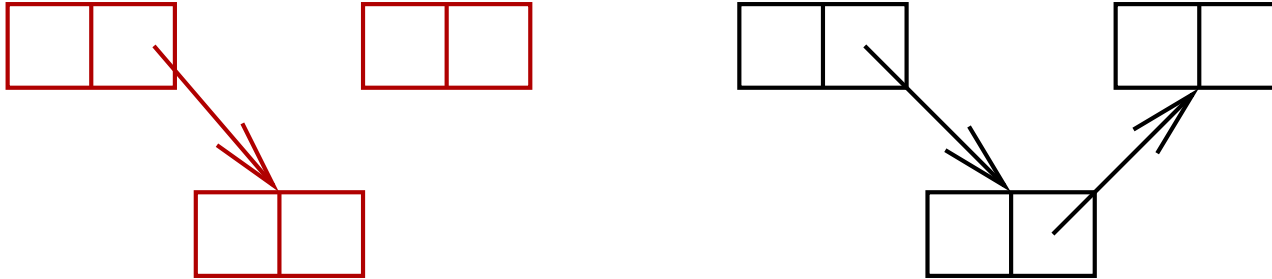
$$a \models \forall R.\phi \text{ iff for all } a' \in a.R^\#, a' \models \phi$$

What does a “must” transition denote?

what must *necessarily* execute:



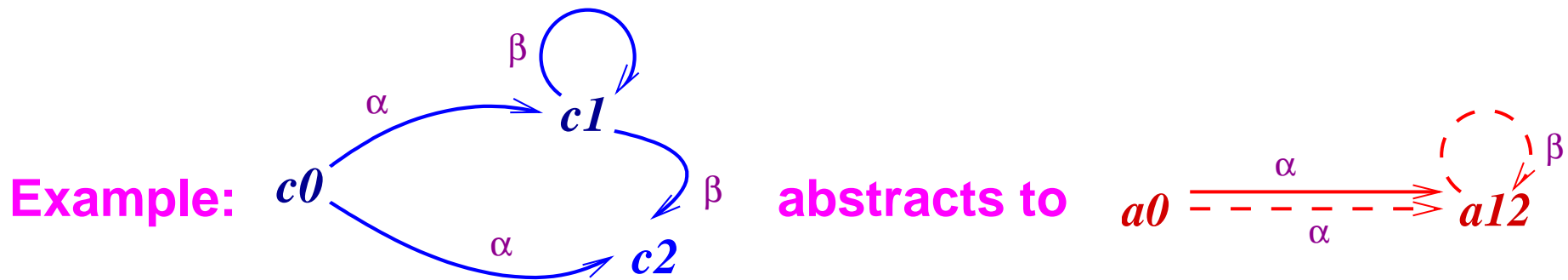
what must *necessarily* be linked:



$a \models \exists R.\phi$ iff there exists $a' \in a.R^b$, $a' \models \phi$

We can validate a full description logic on a MTS

We validate universals on the overapproximated relations and existentials on the underapproximated ones. And we can validate negations via *refutations* on the *dual* approximation.



$$a0 \models^{\text{under}} \exists \alpha. \forall \beta. \neg \text{at_zero}$$

$$\text{iff } a12 \models^{\text{over}} \forall \beta. \neg \text{at_zero}$$

$$\text{iff } a12 \models^{\text{over}} \neg \text{at_zero}$$

$$\text{iff } a12 \not\models^{\text{under}} \text{at_zero}$$

$$\text{iff true}$$

This is fun, but extracting mixed models is hard work! (The reason is that state-abstraction is almost always overapproximating, making it difficult to extract *under*approximating relations.)

Modular validation

Given a model, $\langle \Sigma, \{\tau_\ell\}_{\ell \in L} \rangle$, and $a \in \Sigma$, a judgement $a \models \phi$ is checked with respect to the entire model, that is,

$$a, \langle \Sigma, \{\tau_\ell\}_{\ell \in L} \rangle \models \phi$$

Is there a more “modular” variant of property checking, where a subset of the model is used to validate ϕ ?

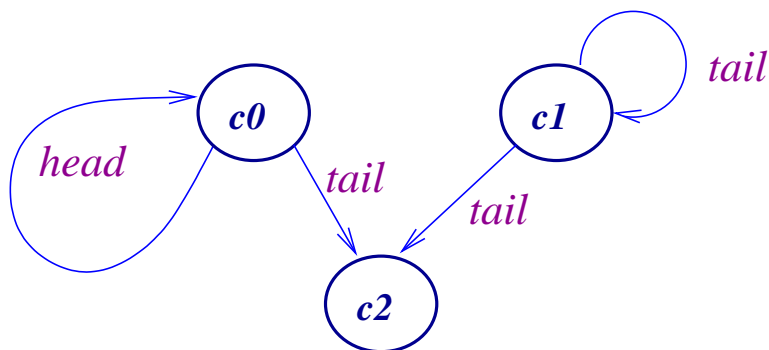
That is, we want to divide the model into disjoint “regions,” h_i , so that we can use this reasoning principle:

$$\frac{h_1 \models \phi_1 \quad h_2 \models \phi_2 \quad h_1 \# h_2}{h_1 \circ h_2 \models \phi_1 * \phi_2}$$

where $h_1 \# h_2$ asserts that h_1 and h_2 are disjoint regions. $\phi_1 * \phi_2$ expresses an assertion whose conjuncts hold true for **disjoint** regions — *no aliasing/sharing of region domains allowed!*

This is the motivation for *separation logic*.

The standard application is to storage heaps, e.g.,
 $h = \langle \text{Cell}, \{\text{head}, \text{tail}\} \rangle$. (Say that $\text{domain}(h) = \text{Cell}$.)



We say that $\langle \text{Cell}_1, \{\text{head}_1, \text{tail}_1\} \rangle \# \langle \text{Cell}_2, \{\text{head}_2, \text{tail}_2\} \rangle$ iff $\text{Cell}_1 \cap \text{Cell}_2 = \{\}$.

The composition of two regions is

$$\langle \text{Cell}_1, \{\text{head}_1, \text{tail}_1\} \rangle \circ \langle \text{Cell}_2, \{\text{head}_2, \text{tail}_2\} \rangle = \\
\langle \text{Cell}_1 \cup \text{Cell}_2, \{\text{head}_1 \cup \text{head}_2, \text{tail}_1 \cup \text{tail}_2\} \rangle$$

For example, the above model can be divided as follows:

$$\langle \{c_0\}, \{ \{(c_0, c_0)\}, \{(c_0, c_2)\} \} \rangle \# \langle \{c_1, c_2\}, \{ \{\}, \{(c_1, c_2), (c_2, c_2)\} \} \rangle$$

The technique can also be used to organize a set of parallel processes.

2. Separation logic (O'Hearn, Pym, Yang)

Additives (the usual classical logic): Let $h \in \text{Heap}$:

$h \models \text{true}$	always
$h \models p \wedge p'$	iff $h \models p$ and $h \models p'$
$h \models \exists x.p_x$	iff exists $v \in \Sigma$ s.t. $h \models p_v$
$h \models E_1 = E_2$	iff $\llbracket E_1 \rrbracket = \llbracket E_2 \rrbracket$

Multiplicatives (based on a commutative partial monoid, $(\text{Heap}, \circ, \epsilon)$):

$h \models \text{emp}$	iff $h = \epsilon$
$h \models p * p'$	iff there exist h_0, h_1 such that $h_0 \# h_1$, $h = h_0 \circ h_1$, $h_0 \models p$, and $h_1 \models p'$
$h \models p -* p'$	iff for all h' , if $h' \# h$ and $h' \models p$, then $h \circ h' \models p'$
$h \models R(E_1, E_2)$... application dependent ...

Separation logic proves correctness properties

Because we can assert disjointness, we can write an assertion that defines a (tail-)noncircular list:

$$\text{nc}(\ell) \text{ iff}_{lfp} (\ell = \text{null}) \vee (\exists c. \text{tl}(\ell, c) * \text{nc}(c))$$

where $h \models \text{tl}(a, b)$ iff $\text{domain}(h) = \{a\}$ and $(a, b) \in \text{tail} \in h$.

The star (*) ensures that *all cells in the list's tail are disjoint from the list's head, addressed by ℓ* . We might prove that a copy function constructs a noncircular list:

```
copy(Cell x ) {  
  nc(x)  
  y := (if x = null  
        then x  
        else new Cell(x.hd, copy(x.tl)));  
  nc(y)  
  return y; }
```

The Reynolds-O'Hearn “small” axioms

$E \doteq E'$ abbreviates $E = E' \wedge emp$ (where emp is the “empty” heap).

$E_1 \mapsto E_2, E_3$ abbreviates $hd(E_1, E_2) \wedge tl(E_1, E_3)$.

Assume that x, a, b , and c are *variables* and that $x \notin \{a, b, c\}$

The small axioms for command forms are

$$\{x \doteq a\} x := E \{x \doteq E[a/x]\}$$

$$\{E \mapsto a, b\} E.tl := E' \{E \mapsto a, E'\}$$

$$\{x \doteq a\} x := \text{new } C(E_1, E_2) \{x \mapsto E_1[a/x], E_2[a/x]\}$$

$$\{x = a \wedge E \mapsto b, c\} x := E.tl \{x = c \wedge E[a/x] \mapsto b, c\}$$

A Hoare triple, $\{\phi\}C\{\psi\}$, defines a relation over commands, *but it also defines “command-transfer functions” that might be used by a data-flow algorithm (e.g., Whaley-Rinard’s analysis, to come).*

Structural rules for inference

The small axioms gain utility when used with these structural rules; let $modified(S)$ be those variables that are targets of assignments in S .

$$\text{Consequence: } \frac{p \supset p' \quad \{p'\}S\{q'\} \quad q' \supset q}{\{p\}S\{q\}}$$

$$\text{Subst: } \frac{\{p\}S\{q\}}{(\{p\}S\{q\})[E_1/x_1, \dots, E_k/x_k]} \quad \text{where } \{x_1, \dots, x_k\} \supseteq free(p, S, q),$$

and $x_i \in modified(S)$ implies E_i is a var, $E_i \notin free(E_j), j \neq i$

$$\text{Frame: } \frac{\{p\}S\{p'\}}{\{p * q\}S\{p' * q\}} \quad \text{where } modified(S) \cap free(q) = \{\}$$

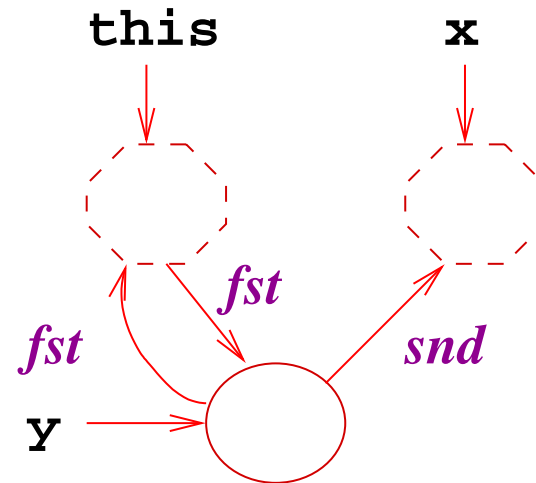
The **Subst** rule motivates the usual rule for procedure invocation (as substitution of actuals for formals).

The **Frame** rule embeds a result proved of a heap “region” into a larger heap, *justifying modular reasoning by reasoning on disjoint heap regions*.

Application of separation logic to Whaley-Rinard analysis

Consider this Whaley-Rinard shape analysis of $m(C\ x)\{\dots\}$:

```
class C {  
  C fst; C snd;  
  function m(C x) {  
    C y := new C(this,x);  
    this.fst := y; }  
}
```

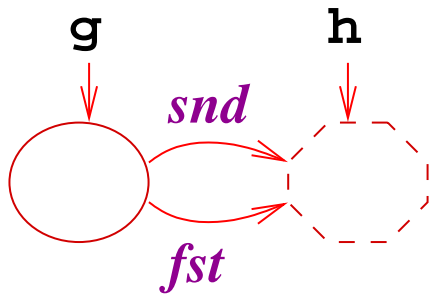


The shape analysis is neatly summarized as this Hoare triple:

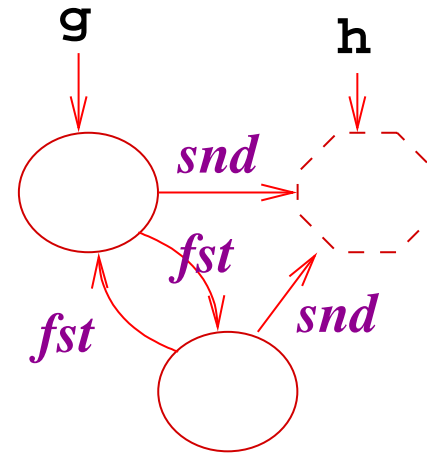
```
m(C x){  
  {this  $\mapsto$  a, b}  
  C y := new C(this,x); this.fst := y;  
  { $\exists y. (this \mapsto y, b) * (y \mapsto this, x)$ }  
}
```

Say there is an invocation, $g.m(h)$:

Caller context:



Result of binding:



The corresponding step in separation logic is

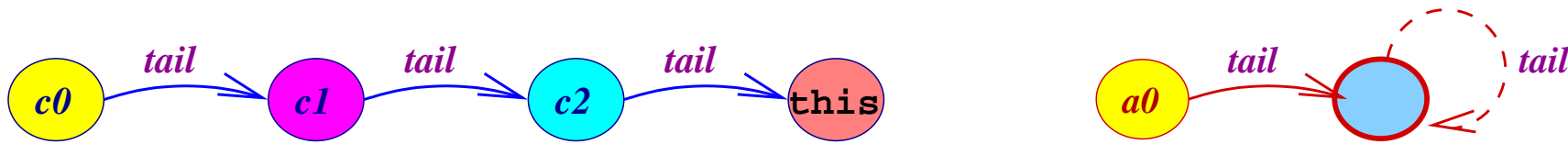
$$\{g \mapsto h, h\} \quad g.m(h) \quad \{\exists y. (g \mapsto y, h) * (y \mapsto g, h)\}$$

because $[this/g, h/x]$ substitutes into the previous triple for $m(C \ x)$:

$$\{this \mapsto a, b\}$$

$C \ y := \text{new } C(\text{this}, x); \text{ this.fst} := y;$

$$\{\exists y. (this \mapsto y, b) * (y \mapsto this, x)\}$$



There are many ad-hoc conventions for asserting noncircularity in shape graphs; separation logic gives us the convention for free.

Consider

```
x := this;
while (...) {
  x := new C(..., x); // \alpha names the summary object
}
```

The loop's approximate invariant, mechanically computed, reads

$$(x = \alpha) \wedge (\alpha \mapsto \dots, (\alpha^* \vee \text{this}))$$

The expression abbreviates a sequence of *distinct objects*:

$$\exists i \geq 0. (x = \alpha_i) \wedge ((\alpha_i \mapsto \dots, \alpha_{i-1}) * (\alpha_{i-1} \mapsto \dots, \alpha_{i-2}) * \dots * (\alpha_0 \mapsto \dots, \text{this}))$$

The Sagiv-Reps-Wilhelm TVLA shape-analysis system uses exactly this interpretation of its summary nodes.

The existential nature of $*$

Recall that \exists -properties cannot be validated on overapproximating models. Nonetheless, separation logic's $*$ *embeds* a useful existential property into its assertions:

Read $h \models \phi_1 * \phi_2$ as *there exists* h_1, h_2 such that $h_1 \# h_2$, $h = h_1 \circ h_2$, $h_1 \models \phi_1$, and $h_2 \models \phi_2$.

The motivating example is

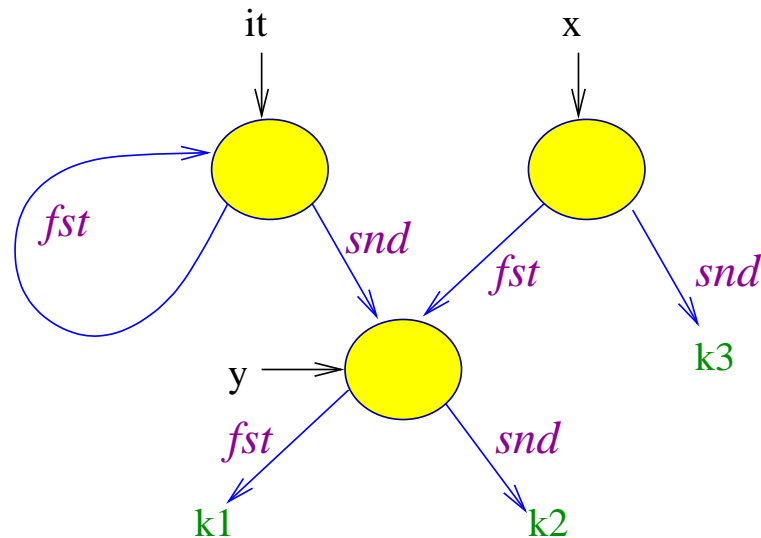
$\text{isAliased}(c)$ iff $tl(\alpha, c) * tl(\beta, c)$

where α and β are placeholders (Prolog-like logical variables) *that must denote distinct cells*.

$\#$ can be checked on abstract values: $\{a\} \# \{a'\}$ iff $a \sqcap a' \neq \perp$, assuming γ is strict. (Or, we can *assume*, as is done in TVLA, that $a \neq a'$ implies $\gamma(a) \cup \gamma(a') = \{\}$.)

“Store-less” models: Path sets

Jonkers and Deutsch proposed “storeless” (heap-less) models:



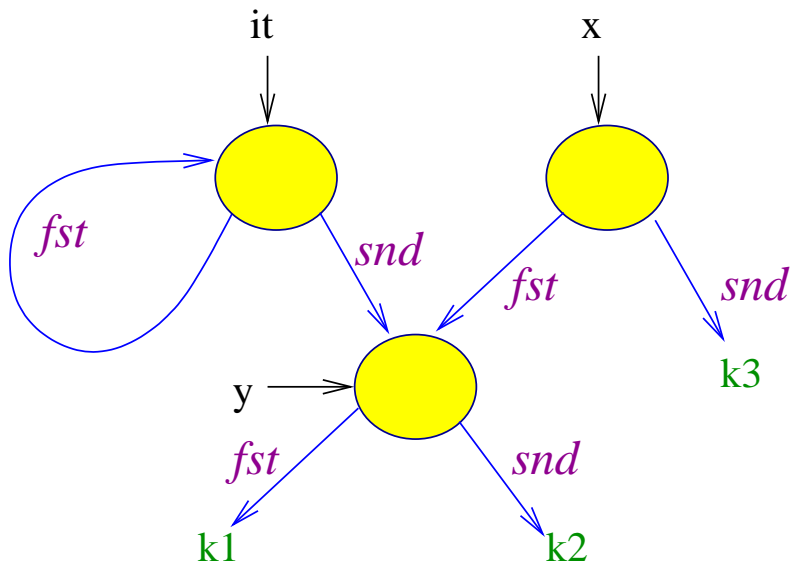
The heap shape is modelled by right-regular equivalence sets of paths from the “entry point,” *it*:

$$\begin{array}{ll} \{fst^i \mid i \geq 0\} & \{fst^i.snd \mid i \geq 0\} \\ \{fst^i.snd.fst \mid i \geq 0\} & \{fst^i.snd^2 \mid i \geq 0\} \end{array}$$

Deutsch developed clever fsa over-approximations of the equivalence classes.

Blanchet's path models

Many questions regarding escapes, leaks, and aliases are answered by the paths from one object of interest to another, e.g., from a global variable to the heap's entry point:



$$\{y.snd^{-1}.fst^i.(fst^{-1})^j \mid i, j \geq 0\}$$

$$\cup \{x.fst.snd^{-1}.fst^i.(fst^{-1})^j \mid i, j \geq 0\}$$

The paths have been normalized by the cancellation law,

$$fst^{-1}.fst \equiv \epsilon$$

The cancellation law gives the paths a pleasant, regular format.

The paths are *traces* through the heap, and questions about the traces can be asked in the language of *linear temporal logic*. Let π be a trace from variable x to it , the result/heap-entry.

We can ask standard questions:

- ◆ Is part of x embedded in the result? $\pi \models at_x \wedge F(des^{-1})$
- ◆ Does x 's cell itself escape in the result? $\pi \models at_x \wedge G(des^{-1})$
- ◆ Is part of x aliased to y ? $\pi \models F(at_y)$
- ◆ Is x a cyclic structure? $\pi \models GF(at_x)$

This approach has been extended by Bazga, Iosif, and Lakhnech.

My recent work

- ◆ Formalizing transition relations as functions of form, $A \rightarrow \mathcal{P}(A)$, where \mathcal{P} is a lower- or upper-powerdomain, and characterizing simulations on such relations in terms of Galois connections on the powerdomains
- ◆ Applying the above to understand the logics that can be validated and refuted on over- and under-approximating models of relations
- ◆ Attempting to integrate separation logic's $*$ operator into modular abstract interpretations of overapproximating models, by studying Blanchet's modular escape analysis for ML.
- ◆ Also, Élodie Sims, a PhD student joint between Kansas and École Polytechnique, has formalized and proved correct an alias (“partitioning”) analysis based on separation logic.

References

This talk: www.cis.ksu.edu/~schmidt/papers

1. B. Blanchet. Ph.D. thesis. <http://www.di.ens.fr/~blanchet>
2. M. Bozga, R. Iosif and Y. Lakhnech. Storeless Semantics and Alias Logic. Proc. PEPM 2003.
3. D. Dams, R. Gerth, O. Grumberg. Abstract Interpretation of Reactive Systems. *ACM TOPLAS* 19 (1997).
4. M. Huth, R. Jagadeesan, D. Schmidt. Modal transition systems: a foundation for three-valued program analysis. Proc. ESOP 2002.
5. C. Loiseaux, et al. Property preserving abstractions for the verification of concurrent systems. *Formal Methods in System Design* 6 (1995).
6. P. O'Hearn, J. Reynolds, H. Yang. Local reasoning about programs that alter data structures. <http://www.dcs.qmul.ac.uk/~ohearn/>
7. D. Pym, P. O'Hearn, H. Yang. Possible worlds and resources: The semantics of BI. *Theoretical Computer Science* <http://www.dcs.qmul.ac.uk/~ohearn/>
8. M. Sagiv, T. Reps, R. Wilhelm. Parametric Shape Analysis via 3-Valued Logic. 26th ACM POPL, 1999.